INSEAD The Business School for the World®

Working Paper

2021/11/TOM

Can AI Help Improve Acute Care Operations? Investigating the Impact of Virtual Triage Technology Adoption

Jiatao Ding
INSEAD, jiatao.ding@insead.edu

Michael Freeman
INSEAD, michael.freeman@insead.edu

Sameer Hasija
INSEAD, Sameer.hasija@insead.edu

To choose the appropriate resources for their healthcare needs (primary care (GP) or emergency department (ED)), patients seeking acute care must self-triage based on their own assessments of symptoms and severity. However, as patients typically lack sufficient medical knowledge, self-triage decisions can often be inaccurate. In response, healthcare and technology companies have been developing and deploying AI-powered virtual triage tools designed to help patients make better self-triage decisions. To date, however, the operational implications of such tools have not been assessed. This paper therefore develops a queueing game model to investigate the impact of virtual triage in the acute care setting and potential policies to maximize its efficacy. We find that, due to its decentralized nature, when virtual triage excessively recommends emergency (primary) care, it counterintuitively brings about a decrease in ED (GP) visits. Another important finding is that in an unregulated environment, the adoption of informative virtual triage can worsen system performance, even when the virtual triage recommendation is reasonably accurate. Building on these insights, we identify two sources of inefficiency and propose associated policy actions that can help unlock the potential operational benefits of virtual triage.

Key words: Healthcare; Acute Care; Virtual Triage; Artificial Intelligence; Digital Operations; Queueing Game
History: Last updated on March 12, 2021.

Can AI Help Improve Acute Care Operations? Investigating the Impact of Virtual Triage Technology Adoption

Jiatao Ding, Michael Freeman, Sameer Hasija
INSEAD, 1 Ayer Rajah Avenue, Singapore 138676
jiatao.ding@insead.edu, michael.freeman@insead.edu, sameer.hasija@insead.edu

To choose the appropriate resources for their healthcare needs (primary care (GP) or emergency department (ED)), patients seeking acute care must self-triage based on their own assessments of symptoms and severity. However, as patients typically lack sufficient medical knowledge, self-triage decisions can often be inaccurate. In response, healthcare and technology companies have been developing and deploying AI-powered virtual triage tools designed to help patients make better self-triage decisions. To date, however, the operational implications of such tools have not been assessed. This paper therefore develops a queueing game model to investigate the impact of virtual triage in the acute care setting and potential policies to maximize its efficacy. We find that, due to its decentralized nature, when virtual triage excessively *recommends* emergency (primary) care, it counterintuitively brings about a *decrease* in ED (GP) visits. Another important finding is that in an unregulated environment, the adoption of *informative* virtual triage can worsen system performance, even when the virtual triage recommendation is reasonably accurate. Building on these insights, we identify two sources of inefficiency and propose associated policy actions that can help unlock the potential operational benefits of virtual triage.

Key words: healthcare, acute care, virtual triage, artificial intelligence, digital operations, queueing game
History: Last updated on March 12, 2021.

1. Introduction

As populations continue to grow and age, acute care services around the world face increasing demand pressure from patients presenting with life-threatening emergencies, acute complications of chronic conditions, and routine illnesses that require prompt attention (Hirshon et al. 2013). Driven by this growth in demand, the value of the global acute care market is expected to expand from USD 2.4 trillion in 2018 to USD 4.0 trillion by 2026, with a compound annual growth rate (CAGR) of 6.7% (Grand View Research 2019). Growing revenue streams along with the increasing volume and diversity of demand have fostered the expansion and comprehensiveness of acute care systems. Consequently, a variety of options have become available to patients to satisfy their acute care needs, including primary and emergency care delivered in varied settings such as primary care practices (also known as general practices (GPs)), hospital-based emergency departments (EDs), and freestanding EDs (Kocher and Ayanian 2016).

This complex range of acute care options creates both opportunities and challenges for patients seeking the appropriate level and location of acute care. Patients with emergency care needs (e.g.,

road traffic accident victims) will typically attend a nearby ED where they can receive timely and prioritized access to experienced emergency practitioners, advanced diagnostic tools, and surgical facilities, if required. To preserve capacity for patients requiring emergency care in EDs, it is normally recommended that patients with less urgent and complex care needs (e.g., patients with flu symptoms or mild pain) instead seek care in a primary care setting, where access to care is typically slower but costs are also lower. However, a patient's choice of the appropriate setting for their healthcare needs critically depends on their ability to self-triage. Yet patients, who lack professional medical knowledge, are charged with determining the urgency and severity of their acute illness or injuries at a time when they may also be experiencing heightened emotions, which can lead to inaccurate disposition decisions (Trivedi et al. 2017).

When patients are unable to self-triage accurately, a mismatch can be created between the supply and demand of acute care resources. On one hand, when patients requiring primary care self-triage wrongly, they may seek care at an ED, incurring unnecessary costs and worsening the *overcrowding problem* at the ED. The arrivals of these patients in EDs are in fact known to be one of the primary drivers of ED overcrowding. For example, studies have shown that the percentage of primary care patients receiving treatment in EDs worldwide ranges from 9% to 60% (Lega and Mengoni 2008). While ED triage (i.e., the prioritization of treatment of patients requiring emergency care, who can preempt other patients if necessary) has been employed as one tool to counteract this problem (Iserson and Moskop 2007), ED treatment of patients who only require primary care nevertheless wastes costly emergency care resources. Moreover, triage provided by nurses and physicians further consumes limited ED resources that might otherwise be used to improve diagnosis and treatment, potentially impacting the quality of care provided (Corl 2019).

On the other hand, a patient requiring emergency care who self-triages wrongly may initially seek care in a primary care setting. If the general practitioner is unable to diagnose or treat the patient, they are then referred to secondary care (e.g., an ED if presenting symptoms are acute, or an outpatient specialist otherwise). In this case, an unnecessary cost is incurred at the GP and patients also experience *treatment delay*. While the prevalence of this problem has received relatively less attention than the prevalence of ED overcrowding, studies have found that patients are generally less likely to underestimate than overestimate their severity (Trivedi et al. 2017). Nevertheless, GPs anecdotally report seeing a non-trivial number of patients who require immediate transfer to an ED (Coyle 2017), with treatment delay having potentially serious and adverse consequences for these patients.

This mismatch in the supply and demand of acute care resources can be alleviated by improving the accuracy of patients' upstream self-triage decisions. As long ago as the 1970s, phone triage services have been used to help patients self-triage (Coons and DuMoulin 2000). When calling a triage

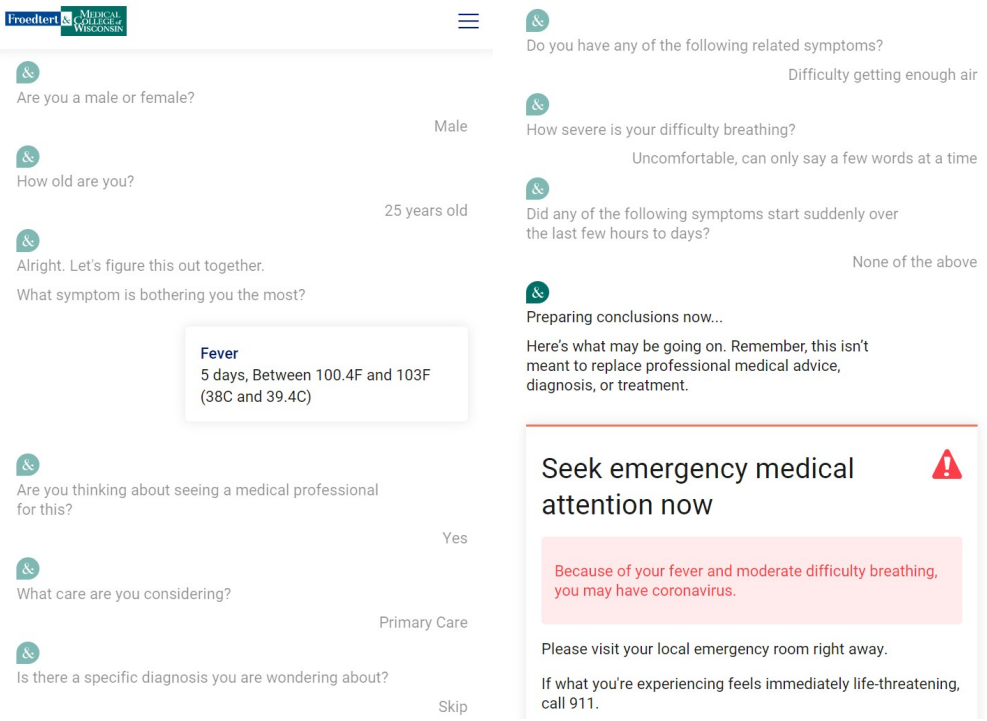
nurse, the patient provides information about their symptoms and receives a triage recommendation. However, phone triage services have longstanding accessibility issues due to the inherent responsiveness-cost trade-off: given high service demand and limited service capability, a call to phone triage typically involves long waiting times, which discourages patients from accessing these services. For example, during the COVID-19 pandemic, people with coronavirus symptoms in the UK struggled to get through on the National Health Service (NHS) 111 phone triage service, with many reporting that they were kept on hold for up to three hours or simply cut off (Dalton 2020).

Recent advances in digital technologies, such as artificial intelligence (AI), offer another solution to improve patient self-triage accuracy and could fundamentally resolve the responsiveness-cost conundrum. In particular, the integration of digital technologies into the management of healthcare operations has been growing steadily, reshaping how care is delivered to patients (Boute and Van Mieghem 2020). Computer-aided diagnosis, telemedicine, wearable medical devices, blockchain-based electronic health records, and computer-assisted drug discovery are a few concrete examples of the digital transformation in healthcare. Meanwhile, to enable more accurate and efficient self-triage, healthcare and tech companies worldwide have been developing and deploying AI-powered virtual triage tools in the form of websites and mobile applications. By asking a sequence of questions relating to patients' personal information and presenting symptoms, virtual triage can give patients immediate triage recommendations before they seek care. As the triage recommendations are provided by pre-trained classification algorithms, virtual triage has a significant cost advantage over traditional phone triage services. Since actual triage nurses are not required, triage service operators find that virtual triage is highly scalable with low marginal operating cost. For patients, virtual triage is highly responsive, meaning they can get instantaneous triage recommendations without delay. Moreover, AI-powered virtual triage enjoys a unique advantage in that triage accuracy improves over time with more training data and better classification algorithms.

Given the potential benefits, virtual triage firms have been partnering with major health providers to increase adoption of the technology. For example, in 2017, Babylon in the UK partnered with the NHS to provide a virtual triage service that, on average, requires 12 text messages and takes about one and a half minutes to complete (Lovett 2018). In the US, Buoy Health offered an AI-powered virtual triage tool to Froedtert & The Medical College of Wisconsin via Froedtert's website (see Figure 1 for the website and a sample virtual triage recommendation). Adoption has been supported by early empirical evidence that has demonstrated the effectiveness of virtual triage in modifying users' care-seeking behavior. For example, a recent study of Buoy Health's virtual triage chatbot showed that 32% of its users reduced their intended level of care (Winn et al. 2019).

Such virtual triage technologies may be especially helpful when patients are suffering with illnesses for which their self-triage accuracy is poor. During the COVID-19 pandemic, for example,

Figure 1 Sample triage recommendation from Buoy Health's AI-powered virtual triage tool on Froedtert's website. Source: https://froedtert.buoyhealth.com/symptom-checker/



patients have struggled to determine the level of care that they require, as both the coronavirus and influenza virus cause respiratory disease that presents as a wide range of similar symptoms. Due to highly infectious nature of the coronavirus, self-triage inaccuracy in this case is extraordinarily dangerous, as it unnecessarily exposes patients and healthcare workers to the risk of cross infection. To ease the burden on healthcare systems and reduce the risk of infection, hospitals and tech firms have developed virtual triage tools to help manage the pandemic (Hao 2020). For example, San Francisco-based tech company GYANT developed the COVID-19 Screener and Emergency Response Assistant, which is a virtual triage tool that has been made accessible to patients on participating hospitals' websites (Blue Shield of California 2020). Furthermore, in China, Tencent built and open-sourced a COVID-19 self-triage assistant with AI technology to "help users with symptoms such as fevers and coughing to conduct a preliminary self-evaluation of their illnesses quickly and seek medical care appropriately" (Tencent 2020).

While virtual triage tools are being developed and deployed worldwide, enabled by advances in AI and expedited by COVID-19, there is at present, however, little understanding of the impact of their use on the healthcare system. In fact, recent efforts to deploy medical AI have found that a lack of understanding of the specific clinical constraints and operational challenges can lead to poor performance in real-world settings, even for medical AI with high accuracy in a lab (Heaven 2020). In particular, multiple studies have shown that virtual triage tends to encourage users to

seek emergency care despite the fact that such care may exceed their needs (Semigran et al. 2015, Chambers et al. 2019). This has invoked widespread concern that the adoption of virtual triage may lead to an increase in ED visits by the so-called "worried-well," thereby worsening the ED overcrowding problem.

With these concerns in mind, this paper develops a queueing game model to explore the operational impact and policy implications of virtual triage adoption by considering a number of practically relevant and related problems. To understand whether excessively recommending emergency care does indeed lead to an increase in ED visits, we first study (1) how virtual triage modifies patient care-seeking behavior. Due to the decentralized nature of the technology, patients may not necessarily follow virtual triage recommendations, particularly when they contradict patients' self-triage decisions. Meanwhile, if and when virtual triage does modify patient care-seeking behavior, one can also ask (2) what is the impact on social cost? Could the adoption of virtual triage lead to a worse outcome with higher equilibrium social cost than before (i.e., in the absence of virtual triage)? If so, (3) how does patient self-triage accuracy moderate the impact of virtual triage? Is the adoption of virtual triage more likely to lead to higher equilibrium social cost when patients have higher or lower self-triage accuracy?

From a policy perspective, our paper also analyzes different policy actions that unlock the operational benefits of virtual triage. It is the unique capability to continuously learn and improve accuracy that underlies the potential benefit of AI-powered virtual triage. However, this advantage of AI also poses a regulatory challenge: (4) is more accurate virtual triage always better? After evaluating a virtual triage tool and deeming it effective, should the regulator limit its authorization to only the current version or also authorize subsequent versions (assumed to have higher accuracy) without re-evaluation (Babic et al. 2019)? Meanwhile, the triage capability of virtual triage can be characterized by the associated receiver operating characteristic (ROC) curve. By varying the discrimination threshold, a particular accuracy (i.e., virtual over-triage and under-triage probability[1]) can be chosen subject to the constraint of the ROC curve. Hence, a question that naturally arises is, (5) for a given virtual triage tool, how should the accuracy, or equivalently, the discrimination threshold, be determined? Moreover, how should the choice of accuracy change as the triage capability of the algorithm improves over time? Lastly, (6) how should the current acute care system respond to the introduction of virtual triage to fully realize its operational benefit?

To answer these questions, in Section 3, a benchmark model is first developed for patients' self-triage and choice of care in the absence of virtual triage. We then present a model of virtual triage in Section 4 and discuss its three unique features: the *accuracy trade-off* as characterized by an ROC curve, the *learning effect* of AI algorithms, and the *cost advantages* of virtual triage technology. In Section 5 and 6, we next investigate the impact of introducing virtual triage into the current acute

care system and show, critically, that naïvely doing so may actually lead to a deterioration in system performance. To resolve the problem, we identify two sources of inefficiency of the equilibrium outcomes in Section 7, i.e., the suboptimality of current GP/ED fees and decisions on virtual triage accuracy subjective to a given ROC curve. We end by proposing policy actions that can help ensure that the operational benefits of virtual triage are realized.

Overall, this paper demonstrates the potential for virtual triage to improve the performance of acute care systems, so long as their implementation and use are properly regulated and the operational implications are carefully considered.

2. Related Literature

Given the focus on acute care, triage, information, and technology adoption, our work contributes to multiple streams of literature relating to healthcare and operations management.

2.1. ED Overcrowding and Triage

The ED overcrowding problem has attracted considerable interest within the operations and healthcare management literature. In this context, a number of papers explore potential triage mechanisms to improve the operational efficiency and responsiveness of EDs. Using a combination of analytical and simulation models, Saghafian et al. (2012) suggest that streaming patients into different groups at the triage stage based on their likelihood of hospital discharge or admission could reduce ED overcrowding in certain situations. Zayas-Cabán et al. (2014) analyze a multi-server two-stage tandem queueing model for a hospital's ED triage and treatment process and identify the optimal control policies. Huang et al. (2012) also study the allocation of physician capacity in EDs, where physicians must choose between serving patients immediately after triage or serving patients whose treatment is already underway. Using a steady-state many server fluid approximation, Kamali et al. (2019) characterize the operational and financial conditions under which provider triage should be applied in addition to traditional nurse triage.

Despite the use of ED triage to prioritize treatment of patients requiring emergency care, the arrival of patients who only require primary care to the ED nevertheless wastes costly resources. ED resources are further stretched as nurses and physicians must provide triage for these patients, while diverting resources away from direct patient care in this way may act to reduce the overall quality of care provided (Corl 2019). In contrast with this traditional ED triage, virtual triage seeks to prevent patients who only require primary care from making unnecessary ED visits in the first place, thus preserving expensive resources for patients with the highest need and reducing the costs associated with providing care that is excessive to patients' needs. Our research contributes to the body of work on triage processes by studying the impact of an upstream decentralized virtual triage service on the ED overcrowding problem.

2.2. Two-Tier Services

An important feature of acute care systems is that patients can typically choose whether to be treated at a GP (tier 1) or an ED (tier 2). As more costly and advanced diagnostic/treatment resources are typically located in the ED, treatment at a GP is normally cheaper but also potentially less effective at resolving a patient's health concerns. Thus, it is typically better for the system if patients requiring primary care visit a GP, thereby reserving expensive tier 2 resources for patients requiring emergency care who stand to benefit from them the most. However, if a patient requiring emergency care arrives at a GP, they will then need to be referred to an ED.

One stream of literature relating to two-tier service considers a system where a tier 1 server (e.g., a generalist) acts as a gatekeeper to a downstream tier 2 server (e.g., a specialist). In these settings, all customers must first be assessed by the tier 1 gatekeeper, who decides whether to serve the customer themselves or, if the customer's service request is too complex, to refer the customer to a downstream specialist. The study of such gatekeeping systems in operations management dates back to Shumsky and Pinker (2003), who derive the optimal referral rate of a tier 1 gatekeeper given a deterministic customer arrival rate and service rate, then analyze the optimal incentive structure in a principal-agent framework. Hasija et al. (2005) extend this framework in a stochastic setting. Lee et al. (2012) analyze a two-tier service system where one or both tiers can be outsourced to a third-party profit-maximizing vendor. Freeman et al. (2017) present an integrated empirical validation of the workload-independence assumption between tier 1 and tier 2 servers, while Freeman et al. (2020) study the accuracy of referral decisions in the ED.

This existing literature focuses on the strategic behavior of service providers, while assuming customers are nonstrategic in the sense that they all initially arrive at a tier 1 server. However, in settings like acute care, depending on their self-assessment of their healthcare requirements, patients can choose to visit a GP first or an ED directly, at their discretion. One recent paper that studies customers' strategic behavior in such a two-tier service setting is Sharma et al. (2019). By modeling patients' choice problem as a network queueing game, they analytically characterize the equilibrium outcomes and design novel incentive mechanisms to align equilibrium patient flow to the social optimum. Building upon their modeling framework, our paper contributes to the two-tier service literature by introducing an additional informative signal from virtual triage and assessing its impact on patient care-seeking behavior and system performance.

2.3. Information in Decentralized Systems

Our paper is closely related to the stream of literature on information in decentralized systems. In the queueing literature, existing work has analyzed how customers' queue-joining behavior depends

on their private information about service quality (Veeraraghavan and Debo 2009, Debo et al. 2012), service rate (Cui and Veeraraghavan 2016), and real-time delay (Hu et al. 2018). Studies within the social learning literature have analyzed firms' pricing (Papanastasiou et al. 2015) and information provision (Papanastasiou et al. 2018) decisions in a context where customers observe the available reviews and update their beliefs regarding product quality. These existing studies on information in decentralized systems focus on *system information* such as product quality, service rate, or queue length, the realizations of which do not vary based on the customers. However, in acute care services, due to the necessity for patients to self-triage before seeking care and to their lack of medical knowledge, customer-specific *personal information* gives rise to a major uncertainty: heterogeneous patients self-triage as one of two types (requiring primary or emergency care) based on their presenting symptoms, with certain probabilities of mis-triage.

Applicable insights from the literature on decentralized systems generally underscore the value of information obfuscation: due to agents' self-interested behavior and the impact of (negative) information externality, full information could lead to suboptimal outcomes, and therefore the optimum is achieved with less or less accurate information. Cui and Veeraraghavan (2016) show that revealing service information to customers may destroy consumer welfare or social welfare. Hu et al. (2018) find that some amount of information heterogeneity in the population can lead to more efficient outcomes than full information. Papanastasiou et al. (2015) illustrate that scarcity strategies can be profitable for a firm when consumers learn according to a quasi-Bayesian rule. Papanastasiou et al. (2018) demonstrate that consumer surplus is nonmonotone in the accuracy of the platform designer's information-provision policy.

While this existing work provides a helpful foundation, since we focus on personal information rather than system information, our findings differ from those of previous studies and are twofold. On one hand, contrary to the existing literature, we find that in our model, full information is strictly preferred because it perfectly reveals a patient's type; they can then seek the appropriate level of care with no uncertainty. On the other hand, we demonstrate that an additional informative signal could either improve or degrade system performance, depending on the signal quality, while a more accurate signal could lead to either better or worse outcomes, depending on the equilibrium regimes. Hence, our findings imply that for the adoption of virtual triage, information obfuscation is preferred conditionally, i.e., only when (more accurate) virtual triage leads to worse outcomes.

2.4. Learning of Personal Information

In diagnostic services, customers are heterogeneous and belong to one of a given set of types. To determine a customer's type, the service provider performs a sequence of imperfect tests. Multiple studies have considered such scenarios. For instance, Alizamir et al. (2013) analyze a congested

system where the diagnostic service provider needs to weigh the benefit of running additional tests to improve diagnosis accuracy against the cost of delaying other customers in the system. In contrast with Alizamir et al. (2013), Sun et al. (2018) examine a scenario where the diagnostic process consists of only a single test, while the subsequent service process is explicitly modeled. Hence, they capture the trade-off between the time spent on diagnosis and time spent on service. Levi et al. (2019) study a similar trade-off where the service provider has to dynamically allocate resources between diagnosing and processing jobs with multiple classes of customers.

In this stream of literature, learning of personal information is centralized, i.e., a single service provider conducts diagnostic tests for arriving customers, and therefore providers face the same information/delay trade-off: while running additional tests could improve diagnosis accuracy, it also delays service for other customers in the system. In such cases, it is clear that more information could be detrimental. However, in our setting, learning of personal information is decentralized, i.e., patients use virtual triage to receive an informative signal about the type of their healthcare needs before seeking care. More importantly, as virtual triage is provided by algorithms, additional information is obtained instantaneously. Hence, in our model, the information/delay trade-off no longer exists and it is initially unclear whether more information could make system performance worse. As we show in later sections, when learning of personal information is decentralized and costless, more information can in fact still degrade system performance.

Our paper also differs from existing work in terms of information control policy. In the existing literature, because of the information/delay trade-off, the objective of service providers is to determine the optimal number of tests to perform. Meanwhile, diagnostic accuracy is assumed to be exogenous. By contrast, in our paper, the diagnostic accuracy can be endogenized subject to a given ROC curve.

2.5. Telemedicine

Our focus on virtual triage is also related to the literature on telemedicine. Rajan et al. (2019) analyze the operational and economic impact of telemedicine technology on a specialist serving a heterogeneous patient population suffering from chronic conditions. Bavafa et al. (2018, 2019) study the impact of telemedicine in a primary care setting. Recent work by Liu et al. (2018) and Savin et al. (2019) analyzes the delivery of telemedicine through on-demand healthcare service platforms. Most similar to our work in this stream of literature is Çakıcı and Mills (2020), who analyze the impact of traditional teletriage provided by nurse-staffed phone lines on healthcare demand management.

Given the focus on AI-powered virtual triage, our study differs from the aforementioned studies in multiple dimensions. First, virtual triage differs from telemedicine in that the technology does

not remotely deliver care to patients. Instead, its main purpose is to assist patients in triaging to the appropriate level of care. Second, virtual triage has a significant cost advantage over traditional nurse line triage. With virtual triage, recommendations are provided by algorithms and no medical professional is required. Third, for patients, virtual triage is more convenient as they can get instantaneous triage recommendations with no delay, compared with traditional nurse lines where long waiting time could be expected given high service demand and limited service capacity. Fourth, the accuracy of virtual triage can be endogenized along the ROC curve and improved over time with more training data and better classification algorithms, while the accuracy of nurse line triage is typically fixed given the training and clinical guidelines the nurses receive. We explore these unique characteristics of virtual triage in this paper.

2.6. Virtual Triage

To the best of our knowledge, the only paper in the literature that also studies virtual triage analytically is Singh et al. (2020). They propose an integral approach where the classifier of virtual triage and/or the queueing system at an ED are jointly optimized to minimize expected waiting cost in the ED. Our paper instead studies virtual triage as a decision support tool for patients who must choose to seek care from a GP or an ED.

Outside of the operations literature, there is a growing interest from the medical community in empirically evaluating the accuracy and effect of virtual triage on users' care-seeking behavior. Verzantvoort et al. (2018) investigate a particular virtual triage smartphone application "Should I see a doctor?" by focusing on app usage, user satisfaction and compliance. Meyer et al. (2020) conduct a cross-sectional survey study of Isabel, an AI-assisted virtual triage tool, and show that a large patient-user group perceives the tool as useful. Semigran et al. (2015) evaluate the triage accuracy of 23 virtual triage tools. They find that triage advice from these tools generally encourages users to seek emergency care. Chambers et al. (2019) review 27 studies on virtual triage and also find that virtual triage tends to recommend emergency care. However, patient compliance with virtual triage in this case is limited: while there is generally good agreement between virtual triage recommendation and patients' intended action, those who the system advises to go to an ED are more likely to seek advice from primary care. This in fact leads to delayed emergency care seeking and a decrease in ED visits. Our paper provides an analytical explanation that reconciles and rationalizes these two seemingly conflicting empirical observations (i.e., for virtual triage to excessively recommend emergency care, but for fewer patients to visit EDs), which is driven by the informativeness of virtual triage recommendations.

3. Modeling Patients' Self-Triage and Choice of Care

In this section, we first establish and analyze a benchmark model in which patients self-triage and choose a level of care (GP or ED) in the absence of virtual triage, building upon the modeling framework from Sharma et al. (2019).

3.1. Model Setting

We consider an acute care system consisting of a number of GPs and an ED serving a patient base. Among the set of patients seeking acute care, some are non-strategic. In particular, patients in highly acute situations or requiring immediate life-saving interventions (e.g., cardiac arrest, major trauma) will visit the ED with certainty, often arriving by ambulance, and will receive prioritized care in the ED. Meanwhile, those patients experiencing a low-acuity illness will, depending on the complexity of their condition, almost certainly visit either a GP or a specialist. These patients typically do not require a same-day acute care appointment and can instead wait for an available appointment on some future date. For these two patient types, we can thus assume that the choice of care location remains unaffected by the introduction of virtual triage technology.

On the other hand, many patients experiencing moderate acute conditions (e.g., chest pain, shortness of breath) are strategic: they have uncertainty regarding the nature of their illness and are therefore unsure whether they should visit a GP first or go directly to the ED. In this case, an additional signal from the virtual triage algorithm about the appropriate location of care for their condition can help reduce their level of self-triage uncertainty and potentially change their care-seeking behavior. The focus of our analysis in this paper is thus on this subset of strategic patients. (Henceforth, we use the terms "strategic patients" and "patients" interchangeably.)

We assume that strategic patients are either of GP-type, denoted by L, or ED-type, denoted by H, and we denote the arrival rate of L patients by λ_L and the arrival rate of H patients by λ_H. While L patients can get treated at either a GP (at a lower cost) or an ED (at a higher cost), H patients require emergency care resources and can only be treated effectively at an ED. Hence, when H patients visit a GP first, they must subsequently be referred to an ED.

Patient self-triage. In deciding whether to visit a GP first or ED directly, strategic patients need to self-triage and determine their type based on their symptoms and medical knowledge. We denote self-triaged GP-type patients by $\hat{L}$, and self-triaged ED-type patients by $\hat{H}$. The arrival rate of $\hat{L}$ patients is denoted by $\lambda_{\hat{L}}$, while the arrival rate of $\hat{H}$ patients is denoted by $\lambda_{\hat{H}}$.

To model self-triage inaccuracy, we assume that a $\hat{L}$ patient has a probability $b_{\hat{L}}$ of being H, while a $\hat{H}$ patient has a probability $1 - b_{\hat{H}}$ of being L. In other words, $b_{\hat{L}}$ and $b_{\hat{H}}$ are the prior beliefs of being ED-type for self-triaged GP-type and self-triaged ED-type patients, respectively.

We assume $b_{\hat{L}} < b_{\hat{H}}$, i.e., $\hat{L}$ patients are less likely than $\hat{H}$ patients to be H. The values of $b_{\hat{L}}$ and $b_{\hat{H}}$ effectively capture the extent of self-triage inaccuracy. When $\hat{L}$ and $\hat{H}$ patients have distinct symptoms, $b_{\hat{L}}$ tends to be small and $b_{\hat{H}}$ tends to be large. On the other hand, when $\hat{L}$ and $\hat{H}$ patients share a wide range of similar symptoms, patients tend to have low self-triage accuracy with $b_{\hat{L}}$ being close to $b_{\hat{H}}$.

Disutility of waiting. We denote strategic patients' disutility of waiting per unit time by w. Unlike non-strategic patients requiring immediate life-saving interventions, who are highly sensitive to delay, or those experiencing routine illness, who are relatively insensitive to delay, the delay sensitivity of strategic patients experiencing moderate acute conditions is mild and similar. In particular, these patients can be thought of as those who require care within 24 hours and are trying to choose between a same-day GP consultation or a trip to the ED. Given this, we assume that both L and H patients have similar delay sensitivity.[2]

Consistent with much of the literature on healthcare system, we denote the expected waiting time at a GP by a constant Q_G (Zorc et al. 2017, Çakıcı and Mills 2020). This follows from the observation that strategic patients seeking acute care only account for a small fraction of all patients accessing GP services. In particular, GPs manage various types of illness, including the delivery of chronic care, treatment of acute non-life-threatening diseases, early detection and referral of patients with urgent serious diseases, health education, immunization, etc. Moreover, to ensure that acute care patients can receive timely prioritized care despite this varied caseload, GPs typically reserve capacity each day for acute care appointments (Gupta and Wang 2008). They also have the ability to reallocate resources between chronic and acute care services and adjust the amount of capacity reserved for acute care appointments, thereby ensuring that waiting times are relatively stable regardless of the arrival rate of strategic patients at a GP, λ_G.

On the other hand, we denote the expected ED waiting time of strategic patients by $Q_E(\lambda_E)$, where $Q_E(\lambda_E)$ is assumed to be strictly increasing and convex in the arrival rate of strategic patients to the ED, λ_E. Unlike a GP, the ED specializes in emergency medicine and is dedicated to acute care. Hence, strategic patients' experiences at EDs, particularly the expected waiting time, critically depend on the care-seeking behaviors of others. From a modeling perspective, the monotonicity and convexity of $Q_E(\lambda_E)$ are satisfied by common queueing models, including $M/M/c$ and $M/G/1$ under a first-in-first-out discipline. Practically, the convexity assumption captures the stochasticity of both the patient arrival process (ED visits are unscheduled, without prior appointments) and treatment process (patients of different characteristics will follow different care pathways) at the ED. The term $wQ_E(\lambda_E)$ thus captures the fact that patients' disutility at the ED increases as more patients visit the ED, due to the growing expected waiting time.

Acute care system operating cost. We assume the expected rates of GP and ED operating costs caused by the arrivals of strategic patients, denoted by $S_G(\lambda_G)$ and $S_E(\lambda_E)$, are increasing and linear in λ_G and λ_E, with $S_G(\lambda_G) = a_G\lambda_G$ and $S_E(\lambda_E) = a_E\lambda_E$. Hence, a_G and a_E denote the expected marginal operating cost per strategic patient arrival to the GP and ED, respectively. Clearly, it is less costly to have an H patient visit the ED directly than visit a GP first, as H patients can only get treated at the ED. Meanwhile, to ensure that it is less costly to have an L patient visit a GP than visit the ED, we assume $a_G + wQ_G \le a_E + wQ_E(\lambda_H)$.

Choice of care. After self-triage, strategic patients compare the expected cost (i.e., the sum of monetary payment and disutility of waiting) of visiting a GP first with the expected cost of going to the ED directly, and they choose the option with a lower cost. Patients incur a monetary payment every time they visit a GP or the ED, denoted by the expected GP fee p_G and expected ED fee p_E. Hence, a $\hat{L}$ patient decides to visit a GP first if $p_G + wQ_G + b_{\hat{L}}[p_E + wQ_E(\lambda_E)] \le p_E + wQ_E(\lambda_E)$ holds, or visit the ED directly otherwise. Similarly, a $\hat{H}$ patient decides to visit a GP first if $p_G + wQ_G + b_{\hat{H}}[p_E + wQ_E(\lambda_E)] \le p_E + wQ_E(\lambda_E)$ holds, or visit the ED directly otherwise.

Let $M_E(\lambda_E)$ denote the marginal cost incurred by an additional ED arrival when the ED arrival rate is λ_E. We therefore have $M_E(\lambda_E) = a_E + wQ_E(\lambda_E) + \lambda_E w\frac{\partial Q_E(\lambda_E)}{\partial \lambda_E}$. The ED overcrowding problem is captured by the term $M_E(\lambda_E)$: when a patient arrives at the ED, in addition to the expected service cost a_E generated, the patient experiences a disutility of waiting $wQ_E(\lambda_E)$ that is caused by the presence of other patients at the ED, while other patients at the ED experience an additional marginal disutility of waiting $\lambda_E w\frac{\partial Q_E(\lambda_E)}{\partial \lambda_E}$ that is caused by the arrival of this patient. We now make the following assumption.

Assumption 1.

(i) $a_G + wQ_G + b_{\hat{L}}M_E(\lambda_H) \le M_E(\lambda_H)$;

(ii) $a_G + wQ_G + b_{\hat{H}}M_E(b_{\hat{L}}\lambda_{\hat{L}} + \lambda_{\hat{H}}) \ge M_E(b_{\hat{L}}\lambda_{\hat{L}} + \lambda_{\hat{H}})$.

Assumption 1 captures the practical conditions of patient self-triage accuracy and acute care cost parameters. Specifically, Assumption 1 implies that, from a social planner's perspective, it is less costly to have $\hat{L}$ patients go to a GP first, as $\hat{L}$ patients can be treated at a GP with a high probability at a much lower cost than at the ED. Meanwhile, it is less costly to have $\hat{H}$ patients visit the ED directly, as $\hat{H}$ patients have a high probability of being H, in which case going to a GP first incurs unnecessary disutility of waiting and service cost, as well as delaying their treatment.

3.2. Social Cost of the Acute Care System

We define social cost $C_s(f_{\hat{L}}, f_{\hat{H}})$ as the sum of strategic patients' disutility of waiting and the service costs of GP and ED operations, where $f_{\hat{L}}$ and $f_{\hat{H}}$ denote the probability of $\hat{L}$ and $\hat{H}$ patients going to ED directly. Then, $C_s(f_{\hat{L}}, f_{\hat{H}})$ can be expressed as follows,[3]

$$C_s(f_{\hat{L}}, f_{\hat{H}}) = \sum_{l\in\{G,E\}} \lambda_l(f_{\hat{L}}, f_{\hat{H}}) w Q_l(\lambda_l(f_{\hat{L}}, f_{\hat{H}})) + S_l(\lambda_l(f_{\hat{L}}, f_{\hat{H}})) \tag{1}$$

where $0 \le f_{\hat{L}}, f_{\hat{H}} \le 1$, $\lambda_G(f_{\hat{L}}, f_{\hat{H}}) = \sum_{\hat{T}\in\{\hat{L},\hat{H}\}}(1-f_{\hat{T}})\lambda_{\hat{T}}$, and $\lambda_E(f_{\hat{L}}, f_{\hat{H}}) = \sum_{\hat{T}\in\{\hat{L},\hat{H}\}}(1-f_{\hat{T}})b_{\hat{T}}\lambda_{\hat{T}} + f_{\hat{T}}\lambda_{\hat{T}}$.

We characterize the social cost function and the associated optimal patient flow by the following proposition.

PROPOSITION 1. *$C_s(f_{\hat{L}}, f_{\hat{H}})$ is jointly convex in $f_{\hat{L}}$ and $f_{\hat{H}}$. In addition, the unique minimum is achieved by $f^*_{\hat{L}} = 0, f^*_{\hat{H}} = 1$.*

In other words, Proposition 1 indicates that in order to achieve social optimum, patients should follow their self-triage decisions despite the uncertainty. This is because, as captured by Assumption 1, having $\hat{L}$ patients at the ED will very likely waste the costly and valuable medical resources and worsen the ED overcrowding problem, while if there are $\hat{H}$ patients at a GP, they tend to incur unnecessary disutility of waiting, service costs and delays in their treatment.

3.3. Incentivizing Optimal Patient Flow

We now study the pricing decisions that induce socially optimal patient flow as determined in Proposition 1. In practice, GP and ED fees cannot be too low because GPs and EDs need to recover operating costs; meanwhile, they cannot be too high due to regulation, competition, and patients' outside options. To capture the joint effect of operating cost, regulation and competition, we focus on the minimum GP and ED fees that can both recover GP/ED operating costs and guarantee optimal patient flow, denoted by $\hat{p}^*_G$ and $\hat{p}^*_E$.[4] We characterize the nonatomic Nash equilibrium (Schmeidler 1973) patient flow, $f^e_{\hat{L}}$ and $f^e_{\hat{H}}$, as well as $\hat{p}^*_G$ and $\hat{p}^*_E$ by the following proposition.[5]

PROPOSITION 2. *For any $p_G, p_E \ge 0$, there exists a unique equilibrium patient flow. In addition, to induce socially optimal patient flow, i.e., $f^e_{\hat{L}} = f^*_{\hat{L}} = 0$ and $f^e_{\hat{H}} = f^*_{\hat{H}} = 1$, GPs and EDs should charge the following fees: $\hat{p}^*_G = a_G, \hat{p}^*_E = a_E + [\lambda_H w \frac{\partial Q_E(\lambda_E)}{\partial \lambda_E}\Big|_{\lambda_H} + wQ_E(\lambda_H) - wQ_E(b_{\hat{L}}\lambda_{\hat{L}} + \lambda_{\hat{H}})]^+$, where $[d]^+ = \max\{0, d\}$.*

Proposition 2 shows that in order to induce optimal patient flow, GPs should charge the expected service cost per patient, while the ED needs to charge a fee that may be higher than the expected

service cost to prevent $\hat{L}$ patients from going to the ED directly. This result is in line with a recent trend of increasing ED fees, which not only recovers operating costs but also prevents people with (self-triaged) non-emergency conditions from visiting EDs "so that a hospital's emergency department can be focused on those who really need emergency services" (Khalik 2014).

3.4. The Incentive Mechanism Conundrum

Due to the inherent self-triage inaccuracy, we still have H patients going to GPs first at rate $b_{\hat{L}}\lambda_{\hat{L}}$ and L patients visiting the ED directly at rate $(1-b_{\hat{H}})\lambda_{\hat{H}}$ even under optimal patient flow. The problem is more salient when patients' self-triage accuracy is poor, that is, $b_{\hat{L}}$ is close to $b_{\hat{H}}$.

If the priority is to further alleviate the ED overcrowding problem, the ED may choose to set a fee higher than $\hat{p}_E^*$ so that even $\hat{H}$ patients have a positive probability of going to a GP first. However, a large number of H patients will then visit a GP first, leading to a potentially serious treatment delay problem. On the other hand, if the priority is to further reduce treatment delay for H patients, the ED may set a fee lower than $\hat{p}_E^*$ so that $\hat{L}$ patients have a positive probability of going to the ED directly. This will lead to a large number of L patients presenting at the ED, worsening the ED overcrowding problem. More importantly, both measures will lead to equilibrium patient flows that deviate from the optimal one characterized by Proposition 1, and will therefore increase social cost.

This conundrum can only be addressed by improving the self-triage accuracy of patients. With the advances in AI technology, virtual triage tools today can provide patients with an immediate triage recommendation before seeking care and so improve self-triage accuracy. In the next section, we present a model of virtual triage.

4. Modeling Virtual Triage

Virtual triage essentially solves a binary classification problem: given the information gathered from the user about their profiles and symptoms, the classification algorithm can triage the user as either GP-type, denoted by $\tilde{L}$, or ED-type, denoted by $\tilde{H}$, and recommend that they visit a GP or an ED accordingly. Specifically, the output of the underlying classification algorithm is a probability s, i.e., the predicted probability of a particular user being H. We assume that s is unbiased. A user with a probability s below (above) a chosen threshold will then be virtual triaged as $\tilde{L}$ ($\tilde{H}$) and recommended to visit a GP (ED).

To characterize the efficacy of the classification algorithm of a given virtual triage tool, let $g(s)$ denote the probability density distribution of the predicted probabilities of all the virtual triage users, with $\int_0^1 sg(s)\,ds = \frac{b_{\hat{L}}\lambda_{\hat{L}}+b_{\hat{H}}\lambda_{\hat{H}}}{\lambda_{\hat{L}}+\lambda_{\hat{H}}}$, i.e., the fraction of H patients in the patient base. We assume $g(s)$ is continuous in $s\in[0,1]$. The function $g(s)$ captures the triage capability of virtual triage.

For instance, if $g(s)$ is distributed towards $s=0$ and $s=1$, virtual triage is highly effective as it can predict the types of users' conditions correctly with a high probability. On the other hand, if $g(s)$ is centered around $s=\frac{b_{\tilde{L}}\lambda_{\tilde{L}}+b_{\tilde{H}}\lambda_{\tilde{H}}}{\lambda_{\tilde{L}}+\lambda_{\tilde{H}}}$, virtual triage is less informative as it cannot distinguish between L and H patients given the input information.

For a given virtual triage tool characterized by the associated $g(s)$, it is then the virtual triage provider's decision to choose a discrimination threshold probability $\bar{s}\in[0,1]$, such that when $s>\bar{s}$, the patient is virtual triaged as $\tilde{H}$ and recommended to visit the ED, and when $s\leq\bar{s}$, the patient is virtual triaged as $\tilde{L}$ and recommended to visit a GP. This threshold $\bar{s}$ is typically chosen with the objective of maximizing or minimizing some scoring function (Gneiting 2011). For instance, the virtual triage provider may assign certain weights to under-triaged and over-triaged cases and then seek to minimize their weighted sum. We next show how the accuracy of virtual triage is determined by the decision on $\bar{s}$.

4.1. The Accuracy Trade-Off of Virtual Triage

For a given $g(s)$, any discrimination threshold probability $\bar{s}$ has an associated (virtual) under-triage probability $\alpha(\bar{s})=Prob(\tilde{L}|H)=\frac{\int_0^{\bar{s}}sg(s)ds}{\int_0^1 sg(s)ds}$ and (virtual) over-triage probability $\beta(\bar{s})=Prob(\tilde{H}|L)=\frac{\int_{\bar{s}}^1(1-s)g(s)ds}{\int_0^1(1-s)g(s)ds}$. As $\bar{s}$ varies, $\alpha(\bar{s})$ and $\beta(\bar{s})$ vary accordingly. (Note that there is a one-to-one mapping between $\bar{s}$ and the pair of α and β. To simplify the exposition, we work with α and β explicitly (and therefore $\bar{s}$ implicitly), and omit the dependence of α and β on $\bar{s}$ for the rest of the paper.) We characterize the implicit dependence of α on β by the following lemma.

LEMMA 1. *The under-triage probability α is a decreasing and convex function in the over-triage probability β, denoted by $\alpha=r(\beta)$, with $r(0)=1, r(1)=0$.*

Lemma 1 highlights the underlying accuracy trade-off faced by the virtual triage provider.[6] When $\bar{s}=0$, virtual triage recommends that all patients seek emergency care, in which case we have under-triage probability $\alpha=0$ and over-triage probability $\beta=1$. As $\bar{s}$ increases, H patients are more likely to be under-triaged as $\tilde{L}$, while L patients are less likely to be over-triaged as $\tilde{H}$. When $\bar{s}=1$, we reach another extreme case where virtual triage recommends that all patients seek primary care, and we have $\alpha=1$ and $\beta=0$. For binary classification models, this accuracy trade-off is commonly captured by a ROC curve, which is defined by plotting sensitivity, i.e., $1-\alpha$, against $1-$specificity, i.e., β, at various thresholds $\bar{s}\in[0,1]$. To facilitate the exposition, rather than working with the ROC curve, we introduce the function $\alpha=r(\beta)$ in Lemma 1 and refer to it as the inverted receiver operating characteristic (IROC) curve.[7]

We now demonstrate the informational effect of virtual triage. For a patient with a prior belief b of being H, let $b_{\tilde{L}}$ denote their posterior belief if the virtual triage recommendation is to see a

GP (i.e., the patient is virtual triaged as $\tilde{L}$). Similarly, let $b_{\tilde{H}}$ denote their posterior belief if the virtual triage recommendation is to visit an ED (i.e., the patient is virtual triaged as $\tilde{H}$). We then have the following corollary.

COROLLARY 1. *For any virtual triage tool, we have* $\alpha \geq 0, \beta \geq 0, \alpha + \beta \leq 1$, *and therefore virtual triage is informative, i.e.,* $b_{\tilde{L}} \leq b \leq b_{\tilde{H}}$.

Corollary 1 highlights the informational benefit of virtual triage: patients are better informed of their type after virtual triage recommendations. In particular, when the virtual triage recommendation is $\tilde{L}$, patients' posterior beliefs of being H are lower than their prior beliefs; when the virtual triage recommendation is $\tilde{H}$, patients' posterior beliefs of being H are higher than their prior beliefs.

4.2. The Learning Effect of Virtual Triage

One unique characteristic of AI-powered virtual triage is its capability to improve accuracy over time with more training data and better learning algorithms. Specifically, $g(s)$ will be distributed more towards $s = 0$ and $s = 1$ over time, achieving a higher triage capability. We formalize the learning effect of virtual triage with the following lemma.

LEMMA 2. *Let* $g_1(s)$ *and* $g_2(s)$ *denote two probability density distributions of the probabilities of being* H *for all the users of the virtual triage tool. Suppose* $\forall\ \bar{s}_1, \bar{s}_2 \in [0,1]$ *s.t.* $\int_{\bar{s}_1}^1 (1-s) g_1(s) ds = \int_{\bar{s}_2}^1 (1-s) g_2(s) ds$, *we have* $\int_0^{\bar{s}_1} s g_1(s) ds \geq \int_0^{\bar{s}_2} s g_2(s) ds$. *Let* $r_1(\beta)$ *and* $r_2(\beta)$ *be the associated IROC curves for* $g_1(s)$ *and* $g_2(s)$. *We then have* $r_2(\beta) \leq r_1(\beta),\ \forall\ \beta \in [0,1]$.

Lemma 2 shows that as the virtual triage tool improves its triage capability, we can have a new IROC curve that lies below the original one. Consequently, we can achieve a higher virtual triage accuracy with lower under-triage and over-triage probabilities. However, as discussed in Babic et al. (2019), this key advantage of AI poses a regulatory challenge: after evaluating a virtual triage tool with a specific pair of under-triage and over-triage probabilities and deeming it effective, should the regulatory authorization be limited to only the current version or also extended to future versions (with presumed higher accuracy) without re-evaluation? To answer this question, we analyze the impact of virtual triage, specifically the learning effect, in Sections 5 and 6.

4.3. The Cost Advantages of Virtual Triage

In addition to the benefits discussed above, another key advantage of virtual triage is the instantaneity of its recommendation: as the virtual triage recommendation is provided by underlying classification algorithms, patients can get instantaneous triage advice with no delay before seeking care. Additionally, virtual triage is highly scalable, with low marginal operating cost and, unlike

traditional phone triage services, does not require triage nurses. An instantaneous, costless and informative (by Corollary 1) virtual triage service may therefore appear to be an obvious win-win for both patients and acute care systems. We investigate this intuition further by studying the operational impact of virtual triage in the next section.

5. The Impact of Virtual Triage on Patient Care-Seeking Behavior

While many virtual triage tools have been implemented, there is no evidence as to how the current acute care system will be affected by or respond to the introduction of virtual triage. Hence, we first study the impact of virtual triage on the current acute care system, as described in Section 3. Note that we assume the virtual triage service is provided free of charge, which reflects the current practice. Moreover, this assumption allows us to focus on understanding the informational effect of virtual triage on patients' care-seeking behavior and social cost.

5.1. Patient Composition

In the virtual triage context, there are four types of patients: $\hat{L}\tilde{L}$, $\hat{L}\tilde{H}$, $\hat{H}\tilde{L}$, and $\hat{H}\tilde{H}$, where the first letter denotes a patient's self-triage decision and the second denotes the virtual triage recommendation.[8] For instance, $\hat{L}\tilde{L}$ patients are those who self-triage as $\hat{L}$ and are virtual triaged as $\tilde{L}$. For a given virtual triage accuracy α and β, patients' posterior probabilities of being H are $b_{\hat{T}\tilde{L}} = \frac{\alpha b_{\hat{T}}}{\alpha b_{\hat{T}} + (1-\beta)(1-b_{\hat{T}})}$ and $b_{\hat{T}\tilde{H}} = \frac{(1-\alpha) b_{\hat{T}}}{(1-\alpha) b_{\hat{T}} + \beta(1-b_{\hat{T}})}$, and the associated arrival rates of each type of patients are $\lambda_{\hat{T}\tilde{L}} = [\alpha b_{\hat{T}} + (1-\beta)(1-b_{\hat{T}})]\lambda_{\hat{T}}$ and $\lambda_{\hat{T}\tilde{H}} = [(1-\alpha) b_{\hat{T}} + \beta(1-b_{\hat{T}})]\lambda_{\hat{T}}$, where $\hat{T} \in \{\hat{L}, \hat{H}\}$. The following lemma characterizes the learning effect of virtual triage on patient composition, i.e., posterior probability and arrival rate, for each type of patients.

Lemma 3.

(i) $\frac{\partial b_{\hat{T}\tilde{L}}}{\partial \alpha} > 0, \frac{\partial b_{\hat{T}\tilde{H}}}{\partial \alpha} < 0, \hat{T} \in \{\hat{L}, \hat{H}\}$.

(ii) $\frac{\partial b_{\hat{T}\tilde{L}}}{\partial \beta} > 0, \frac{\partial b_{\hat{T}\tilde{H}}}{\partial \beta} < 0, \hat{T} \in \{\hat{L}, \hat{H}\}$.

(iii) $\frac{\partial \lambda_{\hat{T}\tilde{L}}}{\partial \alpha} > 0, \frac{\partial \lambda_{\hat{T}\tilde{H}}}{\partial \alpha} < 0, \hat{T} \in \{\hat{L}, \hat{H}\}$.

(iv) $\frac{\partial \lambda_{\hat{T}\tilde{L}}}{\partial \beta} < 0, \frac{\partial \lambda_{\hat{T}\tilde{H}}}{\partial \beta} > 0, \hat{T} \in \{\hat{L}, \hat{H}\}$.

Lemma 3 (i) and (ii) follow from the fact that when the under-triage probability α or over-triage probability β decreases, the virtual triage recommendation is more informative: patients who are virtual triaged as $\tilde{L}$ ($\tilde{H}$) are more likely to be L (H), and therefore their posterior beliefs of being H decrease (increase). Lemma 3 (iii) and (iv) capture the effect of virtual triage accuracy on the arrival rate of each type of patients. Specifically, a decrease in α leads to fewer H patients being under-triaged, reducing the arrival rate of $\tilde{L}$ patients and increasing the arrival rate of $\tilde{H}$ patients. Meanwhile, a decrease in β leads to fewer L patients being over-triaged and therefore increases the arrival rate of $\tilde{L}$ patients and reduces the arrival rate of $\tilde{H}$ patients.

As virtual triage is informative by Corollary 1, we have $b_{\hat{L}\tilde{L}} \leq b_{\hat{L}\tilde{H}}$ and $b_{\hat{H}\tilde{L}} \leq b_{\hat{H}\tilde{H}}$: among patients with the same self-triage decision, those who are virtual triaged as $\tilde{H}$ have a higher posterior probability of being H than those who are virtual triaged as $\tilde{L}$. In addition, as $b_{\hat{L}} < b_{\hat{H}}$ and the posterior probability is monotonically increasing in the prior for any given α and β, we have $b_{\hat{L}\tilde{L}} \leq b_{\hat{H}\tilde{L}}$ and $b_{\hat{L}\tilde{H}} \leq b_{\hat{H}\tilde{H}}$: among patients who receive the same virtual triage recommendation, those who self-triage as $\hat{H}$ have a higher posterior probability of being H than those who self-triage as $\hat{L}$. However, the order of $b_{\hat{L}\tilde{H}}$ and $b_{\hat{H}\tilde{L}}$ is uncertain and depends on the value of α and β. We introduce the following definition to characterize the order of posterior probabilities.

Definition 1 (Triage Dominance - Virtual Triage Accuracy). For a given accuracy of virtual triage, α and β, the accuracy is of *self-triage dominance* if $b_{\hat{L}\tilde{L}} \leq b_{\hat{L}\tilde{H}} < b_{\hat{H}\tilde{L}} \leq b_{\hat{H}\tilde{H}}$; the accuracy is of *virtual triage dominance* if $b_{\hat{L}\tilde{L}} \leq b_{\hat{H}\tilde{L}} \leq b_{\hat{L}\tilde{H}} \leq b_{\hat{H}\tilde{H}}$.

If virtual triage accuracy is of self-triage dominance, patients self-triaged as $\hat{H}$ have higher posterior probabilities than patients self-triaged as $\hat{L}$, regardless of the virtual triage recommendations. On the other hand, if virtual triage accuracy is of virtual triage dominance, patients virtual triaged as $\tilde{H}$ have higher posterior probabilities than patients virtual triaged as $\tilde{L}$, regardless of their self-triage decisions.

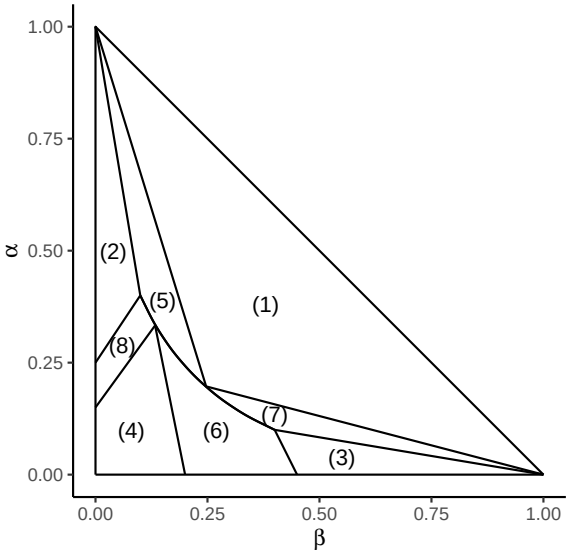
5.2. Patient Flow in Equilibrium

It is important to note that the triage dominance of virtual triage accuracy put forth in Definition 1 does not necessarily imply the dominance of patient care-seeking behavior. Thus, we now analyze the impact of virtual triage on equilibrium patient flow given the expected GP fee $\hat{p}_G^*$ and expected ED fee $\hat{p}_E^*$. Let $f^e_{\hat{L}\tilde{L}}, f^e_{\hat{L}\tilde{H}}, f^e_{\hat{H}\tilde{L}}, f^e_{\hat{H}\tilde{H}}$ denote the probability of patients of each type visiting the ED directly in equilibrium. In particular, we introduce the following definition to characterize the equilibrium patient flow.

Definition 2 (Triage Dominance - Equilibrium Patient Flow). For a given equilibrium patient flow $(f^e_{\hat{L}\tilde{L}}, f^e_{\hat{L}\tilde{H}}, f^e_{\hat{H}\tilde{L}}, f^e_{\hat{H}\tilde{H}})$, it is of *self-triage dominance* if $f^e_{\hat{L}\tilde{L}} = f^e_{\hat{L}\tilde{H}} = 0$ and $f^e_{\hat{H}\tilde{L}} = f^e_{\hat{H}\tilde{H}} = 1$; it is of *virtual triage dominance* if $f^e_{\hat{L}\tilde{L}} = f^e_{\hat{H}\tilde{L}} = 0$ and $f^e_{\hat{L}\tilde{H}} = f^e_{\hat{H}\tilde{H}} = 1$; it is of *GP-type dominance* if $f^e_{\hat{L}\tilde{L}} = f^e_{\hat{L}\tilde{H}} = f^e_{\hat{H}\tilde{L}} = 0$ and $f^e_{\hat{H}\tilde{H}} = 1$; it is of *ED-type dominance* if $f^e_{\hat{L}\tilde{L}} = 0$ and $f^e_{\hat{L}\tilde{H}} = f^e_{\hat{H}\tilde{L}} = f^e_{\hat{H}\tilde{H}} = 1$.

The equilibrium patient flow is of self-triage dominance if patients still follow their self-triage decisions regardless of virtual triage recommendations. Similarly, the equilibrium patient flow is of virtual triage dominance if patients follow virtual triage recommendations and disregard their self-triage decisions. On the other hand, the equilibrium patient flow is of GP-type (ED-type) dominance if patients always go to a GP (ED) unless they are $\hat{H}\tilde{H}$ ($\hat{L}\tilde{L}$). The following proposition proves the uniqueness of equilibrium patient flow and characterizes the equilibrium regimes under different values of α and β.

Figure 2 Eight equilibrium regimes of patient flow.



PROPOSITION 3. *Suppose the expected GP fee is $\hat{p}_G^*$ and expected ED fee is $\hat{p}_E^*$. For any α, β, there exists a unique equilibrium patient flow $(f^e_{\hat{L}\check{L}}, f^e_{\hat{L}\check{H}}, f^e_{\hat{H}\check{L}}, f^e_{\hat{H}\check{H}})$, and we have $f^e_{\hat{L}\check{L}} = 0$ and $f^e_{\hat{H}\check{H}} = 1$ in equilibrium. In addition, depending on the values of α and β, there are eight different equilibrium regimes, characterized by the values of $f^e_{\hat{L}\check{H}}$ and $f^e_{\hat{H}\check{L}}$.*

Pure strategy equilibrium regimes:

(1) $R^e_{0,1} : f^e_{\hat{L}\check{H}} = 0, f^e_{\hat{H}\check{L}} = 1$;

(2) $R^e_{1,1} : f^e_{\hat{L}\check{H}} = 1, f^e_{\hat{H}\check{L}} = 1$;

(3) $R^e_{0,0} : f^e_{\hat{L}\check{H}} = 0, f^e_{\hat{H}\check{L}} = 0$;

(4) $R^e_{1,0} : f^e_{\hat{L}\check{H}} = 1, f^e_{\hat{H}\check{L}} = 0$;

Mixed strategy equilibrium regimes:

(5) $R^e_{(0,1),1} : f^e_{\hat{L}\check{H}} \in (0,1), f^e_{\hat{H}\check{L}} = 1$;

(6) $R^e_{(0,1),0} : f^e_{\hat{L}\check{H}} \in (0,1), f^e_{\hat{H}\check{L}} = 0$;

(7) $R^e_{0,(0,1)} : f^e_{\hat{L}\check{H}} = 0, f^e_{\hat{H}\check{L}} \in (0,1)$;

(8) $R^e_{1,(0,1)} : f^e_{\hat{L}\check{H}} = 1, f^e_{\hat{H}\check{L}} \in (0,1)$.

The relative position of each regime is shown in Figure 2.[9]

First, we observe based on Proposition 3 that when virtual triage recommendations confirm self-triage decisions, patients always follow virtual triage recommendations in equilibrium regardless of their accuracy. In this case, these patients enjoy the informational benefits of virtual triage: while their choices of care locations remain the same as they would have been in the absence of virtual triage, their mis-triage probabilities are now lower as $b_{\hat{L}\check{L}} \le b_{\hat{L}}$ and $b_{\hat{H}\check{H}} \ge b_{\hat{H}}$. On the other hand, when the virtual triage recommendation contradicts the self-triage decision, patients may follow their self-triage decision, the virtual triage recommendation, or a mixed strategy, depending on the values of α and β.

The intuition of the relative positions of the eight equilibrium regimes is as follows. When the accuracy of virtual triage is low, i.e., both the under-triage probability α and over-triage probability β are relatively large, patients' posterior probabilities of being H center around their

prior probabilities. In this case, despite patients' being better informed about their healthcare needs, the equilibrium patient flow is of self-triage dominance and patients still follow their self-triage decisions, resulting in the equilibrium regime $R^e_{0,1}$. If α remains relatively large but β gets smaller, $\hat{L}\tilde{H}$ patients will have a higher posterior probability (i.e., $b_{\hat{L}\tilde{H}}$ increases) and so they will start to visit the ED directly with a positive probability, resulting in the equilibrium regime $R^e_{(0,1),1}$. If β gets very small, $\hat{L}\tilde{H}$ patients will have a posterior probability close to 1 and therefore all of them will follow the virtual triage recommendations instead of their self-triage decisions, resulting in the equilibrium regime $R^e_{1,1}$. Meanwhile, if β instead remains relatively large but α gets smaller, $\hat{H}\tilde{L}$ patients will have a lower posterior probability (i.e., $b_{\hat{H}\tilde{L}}$ decreases) and so they will start to visit a GP first with a positive probability, leading to the equilibrium regimes $R^e_{0,(0,1)}$ and $R^e_{0,0}$.

On the other hand, when both α and β are close to 0, patients simply follow the virtual triage recommendations in equilibrium regardless of their self-triage decisions. This leads to the equilibrium regime $R^e_{1,0}$. If α remains small but β grows larger, $\hat{L}\tilde{H}$ patients will have a lower posterior probability and will start to go to a GP first with a positive probability, resulting in the equilibrium regimes $R^e_{(0,1),0}$ and $R^e_{0,0}$. Meanwhile, if β instead remains relatively small, the increase in α will lead to a higher posterior probability for $\hat{H}\tilde{L}$ patients, who will start to visit the ED directly with a positive probability, resulting in the equilibrium regimes $R^e_{1,(0,1)}$ and $R^e_{1,1}$.

We note that the medical community is increasingly interested in empirically evaluating the accuracy of virtual triage and patients' compliance with virtual triage recommendations. As discussed in Section 2.6, Semigran et al. (2015) and Chambers et al. (2019) have conducted extensive studies and found that virtual triage recommendations tend to encourage patients to seek emergency care. This problem has prompted widespread concern that the adoption of virtual triage could lead to an increase in ED visits, worsening the ED overcrowding problem. However, interestingly, Chambers et al. (2019) also found that while there is generally good agreement between virtual triage recommendations and patients' intended actions, patients who are recommended to go to an ED are more likely to seek primary care. This tendency in fact leads to delayed emergency care seeking and a decrease in ED visits.

Our model provides an explanation that reconciles and rationalizes these two seemingly conflicting empirical findings. When virtual triage excessively recommends emergency care, it leads to high $\lambda_{\hat{L}\tilde{H}}$ and $\lambda_{\hat{H}\tilde{H}}$, with corresponding small value of α and large β, as implied by Lemma 3 (iii) and (iv). Consequently, according to Proposition 3, this will lead to the equilibrium regime $R^e_{0,0}$ (i.e., Region (3) of Figure 2). In this case, notice that there is indeed good agreement between virtual triage recommendations and patients' choice of care: three out of the four types of patients, i.e., $\hat{L}\tilde{L}$, $\hat{H}\tilde{L}$ and $\hat{H}\tilde{H}$ patients, tend to follow virtual triage recommendations. Meanwhile, consistent

with the empirical evidence, the direct arrival rate to the ED actually decreases from $\lambda_{\hat{H}}$ to $\lambda_{\hat{H}\tilde{H}}$, despite virtual triage recommending a high proportion of patients to visit the ED.

These findings highlight a significant way in which virtual triage differs from traditional ED triage: due to its decentralized nature, patients may not necessarily follow virtual triage recommendations. In fact, when virtual triage excessively recommends emergency care, an ED recommendation made by virtual triage carries little information, while a recommendation to see a GP is highly informative. As a result, patients who are recommended to seek primary care will follow the virtual triage recommendation and visit a GP. Meanwhile, patients who are recommended to visit an ED will tend to ignore the recommendations, and many will follow their prior self-triage decisions and instead visit a GP first. Similar arguments hold when virtual triage excessively recommends primary care, which will lead to a decrease in GP visits.

5.3. Learning Effect of Virtual Triage on Patient Strategy in Equilibrium

Next, we study how the learning effect of virtual triage affects patients' strategies in each equilibrium regime. In pure strategy equilibrium regimes, a change of virtual triage accuracy only affects patient composition as characterized by Lemma 3, while the strategies remain the same. In contrast with pure strategy equilibrium regimes, in mixed strategy equilibrium regimes, a change in virtual triage accuracy affects not only patient composition but also the equilibrium strategies for those patients that adopt mixed strategies. In particular, the effect of virtual triage accuracy on patients' mixed strategies in equilibrium is characterized by the following lemma.

Lemma 4.

(i) *In* $R^e_{(0,1),1}$ *and* $R^e_{(0,1),0}$, $\frac{\partial f^e_{\hat{L}\tilde{H}}}{\partial \alpha} < 0, \frac{\partial f^e_{\hat{L}\tilde{H}}}{\partial \beta} < 0.$

(ii) *In* $R^e_{0,(0,1)}$ *and* $R^e_{1,(0,1)}$, $\frac{\partial f^e_{\hat{H}\tilde{L}}}{\partial \alpha} > 0.$

There are two channels through which virtual triage accuracy affects the care choices of patients who adopt mixed strategies. First, virtual triage accuracy has a *direct* effect on their posterior beliefs. Second, it has an *indirect* effect by changing the ED arrival rate of other types of patients who adopt pure strategies, which in turn affects the expected ED waiting time. Lemma 4 follows from considering the impact of these two effects on patient flow.

In particular, in $R^e_{(0,1),1}$ and $R^e_{(0,1),0}$, where a mixed strategy is adopted by $\hat{L}\tilde{H}$ patients, more accurate virtual triage increases their probability of going to the ED directly. As α or β decreases, based on Lemma 3 (i) and (ii), $\hat{L}\tilde{H}$ patients have a higher posterior belief of being H and they are therefore more likely find it beneficial to go to the ED directly. In addition to the higher posterior belief, in $R^e_{(0,1),0}$, lower β also leads to lower $\lambda_{\hat{H}\tilde{H}}$, based on Lemma 3 (iv). This reduces the expected ED waiting time, further increasing the probability of $\hat{L}\tilde{H}$ patients visiting the ED directly. By

contrast, in $R^e_{0,(0,1)}$ and $R^e_{1,(0,1)}$, where $\hat{H}\tilde{L}$ patients adopt a mixed strategy, lower α reduces their probability of going to the ED directly. As α becomes lower, $\hat{H}\tilde{L}$ patients have a lower posterior belief of being H, and they are therefore less likely to go to ED directly. However, while lower β leads to lower posterior belief among $\hat{H}\tilde{L}$ patients, it also reduces $\lambda_{\hat{L}\tilde{H}}$ and $\lambda_{\hat{H}\tilde{H}}$, thereby reducing the expected waiting time at the ED. As a result, lower β could either increase or decrease the probability of $\hat{H}\tilde{L}$ patients visiting the ED directly.

6. The Impact of Virtual Triage on Social Cost

We now analyze the impact of virtual triage on social cost in equilibrium, given an exogenous accuracy α and β.

6.1. Learning Effect of Virtual Triage on Equilibrium Social Cost

Let $C^e_s(\alpha,\beta)$ denote the equilibrium social cost as a function of the virtual triage accuracy under $\hat{p}^*_G$ and $\hat{p}^*_E$. We first characterize how the learning effect of virtual triage affects social cost in equilibrium by the following proposition.

Proposition 4.

(i) *In* $R^e_{0,1}$, *we have* $\frac{\partial C^e_s(\alpha,\beta)}{\partial \alpha} = \frac{\partial C^e_s(\alpha,\beta)}{\partial \beta} = 0$.

(ii) *In* $R^e_{1,1}, R^e_{0,0}$ *and* $R^e_{1,0}$, *we have* $\frac{\partial C^e_s(\alpha,\beta)}{\partial \alpha} > 0, \frac{\partial C^e_s(\alpha,\beta)}{\partial \beta} > 0$.

(iii) *In* $R^e_{(0,1),1}$ *and* $R^e_{(0,1),0}$, *we have* $\frac{\partial C^e_s(\alpha,\beta)}{\partial \beta} < 0$.

(iv) *In* $R^e_{0,(0,1)}$ *and* $R^e_{1,(0,1)}$, *we have* $\frac{\partial C^e_s(\alpha,\beta)}{\partial \alpha} > 0$.

Proposition 4 (i) directly follows from the fact that patients follow their self-triage decisions in $R^e_{0,1}$. For Proposition 4 (ii), in pure strategy equilibrium regimes $R^e_{1,1}, R^e_{0,0}$ and $R^e_{1,0}$, patients follow either their self-triage decisions or virtual triage recommendations with certainty. In this case, lower α reduces λ_G, while λ_E is independent of α (as patients who are under-triaged will go to the ED regardless, either directly or referred by a GP). Hence $C^e_s(\alpha,\beta)$ decreases with lower α. On the other hand, if β is lower, fewer patients will be over-triaged and go to the ED directly. This will reduce the total ED arrival rate and increase the GP arrival rate by the same amount. Since a patient visit to a GP is less costly than a visit to the ED, $C^e_s(\alpha,\beta)$ decreases with lower β.

The learning effect of virtual triage on equilibrium social cost for mixed strategy equilibrium regimes is more complicated. The key implication highlighted by Proposition 4 (iii) and (iv) is patients' inability to internalize the information externality of their behavior under mixed strategy equilibrium regimes, which dominates the change of equilibrium social cost. In particular, when $\hat{L}\tilde{H}$ patients adopt mixed strategies, lower β will increase their posterior belief that they are H, and they are therefore more likely to go directly to the ED. However, their behavior does not internalize

the *negative* information externality (i.e., longer waiting times) exerted on other patients at the ED. As a result, the negative information externality dominates the changes in equilibrium social cost, and therefore $C_s^e(\alpha,\beta)$ increases with lower β in $R^e_{(0,1),1}$ and $R^e_{(0,1),0}$. On the other hand, when $\hat{H}\tilde{L}$ patients adopt mixed strategies, lower α will decrease their posterior belief that they are H, and they are therefore less likely to go directly to the ED. Similarly, their behavior does not internalize the *positive* information externality (i.e., shorter waiting times) exerted on other patients at the ED. As a result, $C_s^e(\alpha,\beta)$ decreases with lower α in $R^e_{0,(0,1)}$ and $R^e_{1,(0,1)}$.

6.2. Benchmarking the Impact of Virtual Triage on Equilibrium Social Cost

Since equilibrium social cost is nonmonotone in virtual triage accuracy as shown by Proposition 4, an important question follows: can the adoption of virtual triage lead to an equilibrium social cost that is higher than before (i.e., in the absence of virtual triage), even though virtual triage is informative? The following proposition establishes the existence of such equilibria.

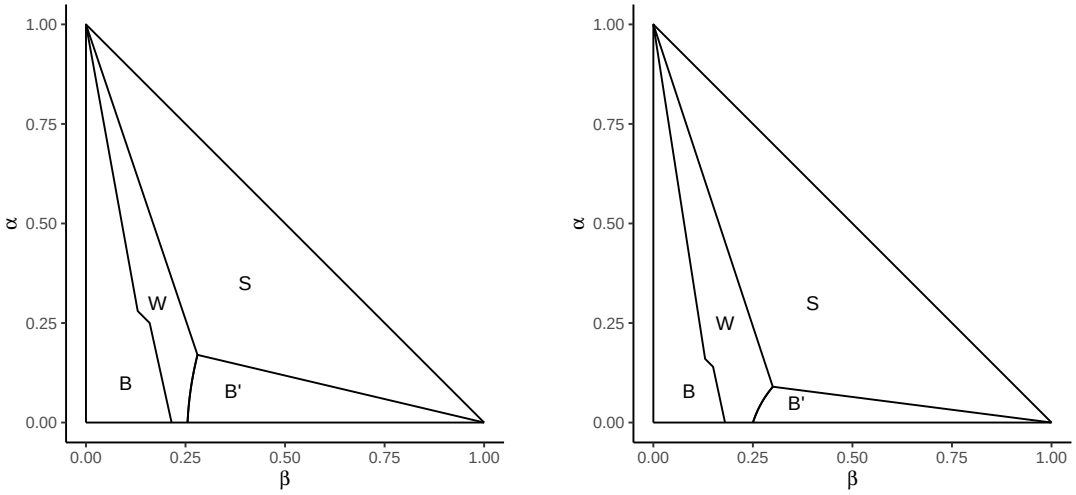
PROPOSITION 5. *There exists* α,β *s.t.* $C_s^e(\alpha,\beta) > C_s(f_{\hat{L}}=0, f_{\tilde{H}}=1)$.

Given the results showing that (1) virtual triage could lead to a higher equilibrium social cost, by Proposition 5, (2) lower β leads to higher equilibrium social cost when $\hat{L}\tilde{H}$ patients adopt mixed strategies, by Proposition 4, and (3) the relative positions of equilibrium regimes $R^e_{(0,1),1}$ and $R^e_{(0,1),0}$ as shown in Figure 2, we would expect equilibrium outcomes that are worse than before to be achieved under relatively small (but not especially small) values of β. Figure 3 visualizes the impact of virtual triage on equilibrium social cost.[10] Region S is where virtual triage has no impact on equilibrium social cost; Regions B and B′ are where virtual triage reduces equilibrium social cost; and Region W is where virtual triage leads to higher equilibrium social cost.

When α is relatively large and β is relatively small, we have the upper Region W: $\hat{L}\tilde{H}$ patients have relatively but not especially high $b_{\hat{L}\tilde{H}}$, and in equilibrium, many of them find it beneficial to go to the ED directly. Each such visit will, with a high probability, waste costly ED resources and generate a negative externality on other patients at the ED, therefore leading to a higher equilibrium social cost. Meanwhile, as α becomes smaller and β remains relatively small, we have the bottom Region W: although $b_{\hat{L}\tilde{H}}$ gets higher with smaller α by Lemma 3 (i), $\lambda_{\hat{L}\tilde{H}}$ also gets higher by Lemma 3 (iii). As a result, while each direct visit to the ED from $\hat{L}\tilde{H}$ patients is less likely to waste ED resources and generate a negative externality on other ED patients, the volume effect of higher $\lambda_{\hat{L}\tilde{H}}$ could still lead to a higher equilibrium social cost. We notice that even with $\alpha=0$, i.e., all H patients are perfectly revealed and therefore the treatment delay problem at GPs is eliminated, a relatively small β could still lead to a worse equilibrium outcome.

There are two separate regions (Regions B and B′) where equilibrium social cost is lower than it was in the absence of virtual triage. Region B′ is driven by the alleviation of the ED overcrowding

Figure 3 Impact of virtual triage on equilibrium social cost. Region S is where virtual triage has no impact on equilibrium social cost; Regions B and B′ are where virtual triage reduces equilibrium social cost; Region W is where virtual triage increases equilibrium social cost.



(a) Low self-triage accuracy.

(b) High self-triage accuracy.

problem: with small α and large β, the posteriors of $\hat{H}\tilde{L}$ patients are considerably lower than their priors, while the posteriors of $\hat{L}\tilde{H}$ patients are close to their priors. As a result, $\hat{H}\tilde{L}$ patients will go to a GP first with a positive probability in equilibrium, while $\hat{L}\tilde{H}$ patients still follow their self-triage decisions. This will reduce direct arrivals to the ED, thereby alleviating the overcrowding problem. On the other hand, Region B is primarily driven by the alleviation of the treatment delay problem at GPs: with sufficiently small β, $b_{\hat{L}\tilde{H}}$ is sufficiently high and all $\hat{L}\tilde{H}$ patients go to the ED directly in equilibrium. Consequently, the benefit of alleviating the potential treatment delay problem for $\hat{L}\tilde{H}$ patients outweighs the negative externality at the ED. Moreover, if both α and β are sufficiently small, patients simply disregard their self-triage decisions and follow the virtual triage recommendations in equilibrium. Hence, we could achieve an equilibrium patient flow that alleviates both treatment delay at GPs and ED overcrowding, which corresponds to the bottom of Region B.

We provide further insight into how the impact of virtual triage on the equilibrium social cost is moderated by self-triage accuracy. When patient self-triage accuracy is lower, the optimal patient flow in the absence of virtual triage suffers from a higher level of inefficiency: more ED-type patients visit a GP first and more GP-type patients go to the ED directly. In this case, less accurate virtual triage should be able to achieve an equilibrium social cost that is lower than before, resulting in Regions B and B′ having a larger area. By comparing Figure 3 (a) and (b), we can see that when patient self-triage accuracy is lower, Regions B and B′ do indeed have a larger area.

These findings also shed light on our discussion of the regulatory challenges surrounding medical AI, particularly given AI's ability to become more accurate over time. We show that, in an unregulated environment, the adoption of informative virtual triage with reasonably high accuracy could still lead to a deterioration in system performance if the scoring function and the associated virtual triage accuracy are chosen naïvely. Furthermore, as the accuracy of virtual triage increases, so too might the equilibrium social cost. Consequently, updating a virtual triage algorithm to a more accurate version without re-evaluation and additional regulatory approval may actually reverse any previously demonstrated benefit.

7. Unlocking the Operational Benefits of Virtual Triage

The inefficiency of equilibrium outcomes as characterized in Sections 5 and 6 has two sources. First, for the *healthcare provider*, the current GP and ED fees, which induce optimal patient flow in the absence of virtual triage, may be suboptimal after the adoption of virtual triage. Second, for the *virtual triage technology provider*, their chosen scoring function and decision on virtual triage accuracy subject to a given IROC curve may be suboptimal in terms of the minimization of social cost (Gneiting 2011). In this section, then, we explore associated policy actions to enable healthcare systems to reap the operational benefit of virtual triage.

7.1. Optimizing Over Patient Flow Under Exogenous Virtual Triage Accuracy

For a given virtual triage accuracy of α and β, let $f^*_{\hat{L}\tilde{L}}, f^*_{\hat{L}\tilde{H}}, f^*_{\hat{H}\tilde{L}}, f^*_{\hat{H}\tilde{H}}$ denote the optimal patient flow and $C^*_s(\alpha,\beta)$ denote the associated minimum social cost. We then have the following proposition.

PROPOSITION 6. *For any α, β, we have $f^*_{\hat{L}\tilde{L}} = 0, f^*_{\hat{H}\tilde{H}} = 1$. In addition, we have $\frac{\partial C^*_s(\alpha,\beta)}{\partial \alpha} \geq 0$ and $\frac{\partial C^*_s(\alpha,\beta)}{\partial \beta} \geq 0$, and therefore $C^*_s(\alpha,\beta) \leq C_s(f_{\hat{L}} = 0, f_{\hat{H}} = 1)$.*

Two points of Proposition 6 are worth discussing. First, we always have $f^e_{\hat{L}\tilde{L}} = f^*_{\hat{L}\tilde{L}} = 0$ and $f^e_{\hat{H}\tilde{H}} = f^*_{\hat{H}\tilde{H}} = 1$ under both equilibrium patient flow and optimal patient flow. This observation highlights the effectiveness of virtual triage when it confirms patients' self-triage decision under equilibrium patient flow. Hence, the inefficiency of equilibrium patient flow is due to those cases where the virtual triage recommendations contradict patients' self-triage decisions. In the case of these patients, the current acute care system (with expected GP fee $\hat{p}^*_G$ and ED fee $\hat{p}^*_E$) may fail to incentivize them to behave optimally. Second, we notice that, not surprisingly, more accurate virtual triage always leads to lower social cost under coordinated optimal patient flow. In this case, the informational benefit of virtual triage is fully realized.

GP and ED fees may therefore need to be adjusted to induce optimal patient flow after the adoption of virtual triage. The minimum GP and ED fees that can both recover acute care system operating cost and induce optimal patient flow, which we denote by $\tilde{p}^*_G(\alpha,\beta)$ and $\tilde{p}^*_E(\alpha,\beta)$, are characterized by the following proposition.

PROPOSITION 7. *For any α, β, we have $\tilde{p}_G^*(\alpha,\beta) = \hat{p}_G^*$. In addition, for any α, β s.t. $f^*_{\hat{L}\tilde{H}} \in (0,1)$ or $f^*_{\hat{H}\tilde{L}} \in (0,1)$, we have $\tilde{p}_E^*(\alpha,\beta) > \hat{p}_E^*$.*

Proposition 7 shows that in order to induce optimal patient flow, the ED fee may need to be increased while the GP fee remains unchanged. However, adjusting the ED fee to induce optimal patient flow has two major downsides. First, given virtual triage's capability to improve in accuracy over time, the ED fee will have to be adjusted dynamically, which may be difficult to achieve in practice due to regulation and competition. In addition, a higher ED fee will further add to the cost burden of patients seeking emergency care.

7.2. Optimizing Over IROC Curves Under Current GP and ED Fees

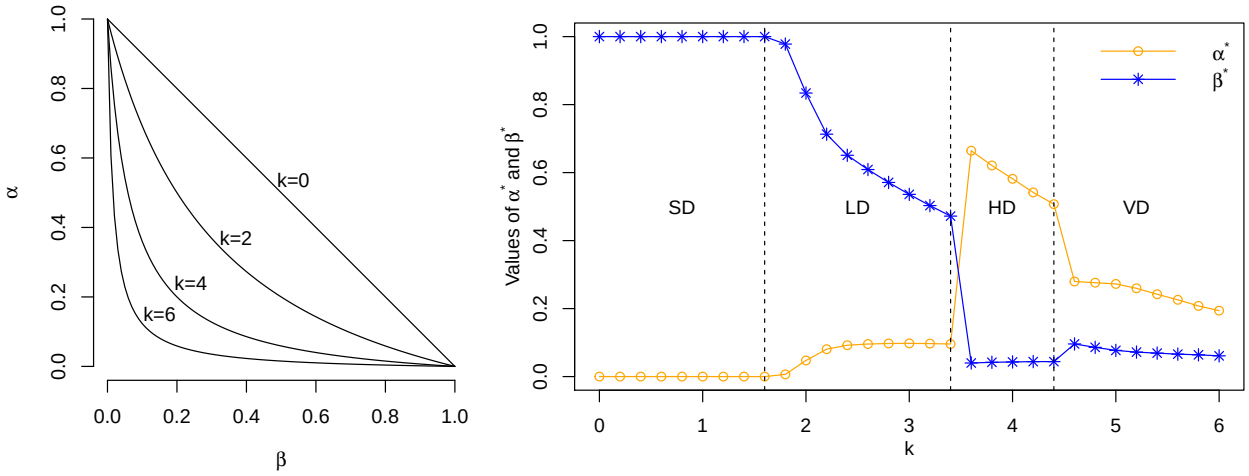
Given the potential limitations of dynamic fee adjustment in this context, we now study the second source of inefficiency, which is the virtual triage provider's decision regarding accuracy. Note that the accuracy decision is subject to the constraint that $\alpha = r(\beta), \beta \in [0,1]$, while we assume that the GP and ED fees remain unchanged at $\hat{p}_G^*$ and $\hat{p}_E^*$, respectively.

Clearly, compared with the minimum social cost in the absence of virtual triage, equilibrium social cost after the adoption of virtual triage will not be increased by endogenizing β: $\bar{s}$ can be set to either 0 or 1, in which case we have $r(\beta) + \beta = 1$, and patients' posteriors remain the same as their priors. More importantly, we would like to minimize the equilibrium social cost. This is equivalent to using the equilibrium social cost as the scoring function to determine the optimal over-triage probability β^* and under-triage probability $\alpha^* = r(\beta^*)$ for a given IROC curve. As there are multiple equilibrium regimes, as shown by Proposition 3, and as the IROC curve can intersect with different equilibrium regimes, explicit analytical characterization of α^* and β^* is infeasible. Hence, we perform numerical analysis to study this problem.

For the numerical analysis, we assume the IROC curve $\alpha = r(\beta)$ takes the implicit functional form $(1-\alpha)(1-\beta)2^{-k} = \alpha\beta, k \in [0,\infty)$. It is easy to verify that α is a decreasing and convex function of β, with $r(0) = 1$ and $r(1) = 0$. The parameter k effectively captures the triage capability of virtual triage, as shown in Figure 4 (left). As k increases, a higher triage capability is achieved.

Figure 4 (right) shows the optimal virtual triage accuracy α^* and β^* that minimize the equilibrium social cost, and how the optimal accuracy changes as k increases.[11] We notice that there are four regions characterized by the equilibrium patient flow: equilibrium patient flow of self-triage dominance (SD), GP-type dominance (LD), ED-type dominance (HD), and virtual-triage dominance (VD). When k is small, the triage capability of the virtual triage algorithm remains low. Therefore, despite being better informed of the true nature of their healthcare needs after using the technology, patients' posteriors still center around their priors. Consequently, for any

Figure 4 (Left) IROC curves associated with virtual triage algorithms with different triage capabilities, and (right) the corresponding optimal β^* and $\alpha^* = r(\beta^*)$ to minimize equilibrium social cost under expected GP fee $\hat{p}_G^*$ and expected ED fee $\hat{p}_E^*$.



$\beta \in [0,1], \alpha = r(\beta)$, patients still follow their self-triage decisions in equilibrium. Hence, any virtual triage accuracy along the IROC curve is in fact optimal and will induce an equilibrium patient flow of self-triage dominance. Without loss of generality, we plot $\alpha^* = 0, \beta^* = 1$ in Region SD.

As k increases and the triage capability of virtual triage becomes higher, equilibrium patient flow of lower social cost can be achieved with a sufficiently small α^* at the cost of relatively large β^*. In particular, the optimum is achieved with both $\hat{L}\tilde{H}$ and $\hat{H}\tilde{L}$ patients going to a GP first, thereby generating an equilibrium patient flow of GP-type dominance which alleviates the ED overcrowding problem. As k further increases, the optimum is achieved under an equilibrium patient flow of ED-type dominance, which requires a sufficiently small β^* at the cost of relatively large α^*. Finally, as the triage capability of virtual triage technology becomes sufficiently high, the optimum is achieved under an equilibrium patient flow of virtual triage dominance, with both small α^* and small β^*.

We note that in Region LD, as k increases, α^* increases while β^* decreases. This is despite the fact that with larger k there exists virtual triage accuracy along the IROC curve for which both α and β can be reduced. The increase in α^* can be explained, however, by observing that as k increases, the IROC curve becomes flatter in regions with large β. Consequently, it is possible to shift along the IROC curve to a new point with a significantly smaller β^* at the cost of only a slight increase in α^*. In other words, by slightly worsening the treatment delay problem, the ED overcrowding problem can be significantly alleviated. Similarly, in Region HD where optimal virtual triage accuracy is achieved with small β^*, as k increases, it is possible to reduce α^* significantly with a slight increase in β^*. Hence, by slightly increasing ED overcrowding, the treatment delay problem at GPs can be significantly alleviated. We also observe in Figure 4 (right) that the rate of increase of β^* with respect to k in Region HD is smaller than the rate of increase of α^* in Region

LD. This can explained by the large negative information externality at the ED, meaning that the cost of a larger over-triage probability in Region HD is higher than the cost of a larger under-triage probability in Region LD.

In contrast with what we observe in Regions LD and HD, in Region VD we see that when the triage capability of the virtual triage tool becomes sufficiently high, optimal accuracy is achieved with both small α^* and small β^*. In this region, as k increases, both α^* and β^* are reduced under optimal virtual triage accuracy. In other words, the treatment delay problem at GPs and the ED overcrowding problem can be jointly alleviated as the algorithm continues to improve. As k approaches ∞ in Region VD, both α^* and β^* decrease monotonically and approach 0.

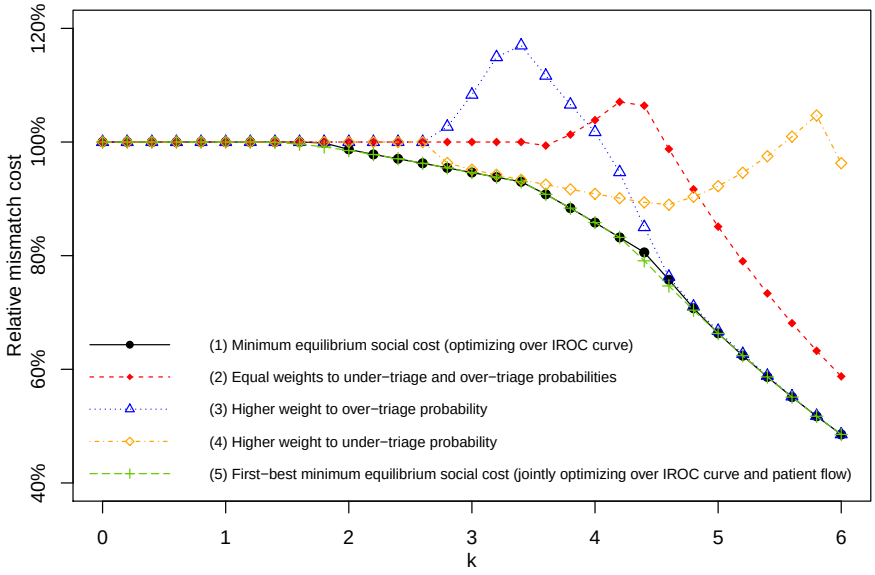
Overall, our results show that, critically, the optimal decision regarding virtual triage accuracy subjective to a given IROC curve is nuanced by and nonmonotone in its triage capability. Specifically, the optimal under-triage or over-triage probability may increase as the triage capability improves. Moreover, when the triage capability is not sufficiently high, it may be optimal to have rather large under-triage or over-triage probability.

7.3. Mismatch Costs Under Different Scoring Functions

Next, we consider the impact of different scoring functions chosen by the virtual triage provider on equilibrium social cost. To characterize the inefficiency of equilibrium patient flow, we define the *mismatch cost* as the increase in social cost that results from L patients going to the ED and H patients visiting a GP. We denote this mismatch cost under optimal patient flow in the absence of virtual triage by C^m. After the introduction of virtual triage, we consider how this mismatch cost changes under four alternative scoring functions. The first seeks to minimize the equilibrium social cost by optimizing over the IROC curve under current GP and ED fees (i.e., as described in Section 7.2 and in accordance with Figure 4 (right)). The second to the forth assign certain weights to under-triage and over-triage probabilities and seek to minimize the weighted sum of both. In particular, the second assigns equal weights to under-triage and over-triage probabilities, i.e., $\beta = r(\beta)$; the third (forth) assigns a higher weight of 0.75 to the over-triage (under-triage) probability, with a weight of 0.25 assigned to the under-triage (over-triage) probability. For these four scoring functions, let $C_k^{m*}, C_k^{m=}, C_k^{m<}$ and $C_k^{m>}$ denote the mismatch cost associated with a triage capability k under current GP and ED fees, respectively. Line (1) to Line (4) in Figure 5 plot $C_k^{m*}/C^m, C_k^{m=}/C^m, C_k^{m<}/C^m$, and $C_k^{m>}/C^m$, respectively, i.e., the relative mismatch cost under these four scoring functions.

Line (1) in Figure 5 shows that when the triage capability is relatively low, the minimum equilibrium social cost from optimizing over the IROC curve alone remains the same as in the absence of virtual triage. As k increases, we can achieve a minimum equilibrium social cost that is lower, with

Figure 5 The relative mismatch cost associated with equilibrium patient flow after the introduction of virtual triage. Line (1) has a scoring function which minimizes equilibrium social cost by optimizing over the IROC curve under current GP and ED fees; Lines (2), (3) and (4) use a scoring function which, under current GP and ED fees, assigns weights of 0.5, 0.75 and 0.25 to the over-triage probability and 0.5, 0.25 and 0.75 to the under-triage probability, respectively; Line (5) has the same scoring function as Line (1), but jointly optimizes over both the IROC curve and patient flow.



the relative mismatch cost decreasing monotonically as k increases. Line (2) to Line (4) highlight the problem with simple scoring functions, i.e., naïvely assigning fixed weights to the under-triage and over-triage probabilities. We can see that to achieve a relative mismatch cost lower than 100%, these scoring functions will require a virtual triage algorithm of a higher triage capability (i.e., with a larger k). More importantly, due to the nonmonotonicity of the equilibrium social cost in virtual triage accuracy (as shown by Proposition 4), a higher triage capability can actually increase the mismatch cost when a scoring function is chosen such that it does not have the objective of minimizing the equilibrium social cost. Lastly, note that when the triage capability of the virtual triage technology is sufficiently high, even a naïvely set scoring function can reduce the mismatch cost: the gap between Line (1) and Line (2) to Line (4) will decrease and approach to 0 as k increases and approaches ∞.

We further examine a variant on the first scoring function (which minimizes the equilibrium social cost), where we not only optimize over the IROC curve but also patient flow (by potentially adjusting the ED fee, as characterized by Proposition 7). Letting C_k^{m**} denote the associated mismatch cost, Line (5) in Figure 5 plots the relative mismatch cost (i.e., C_k^{m**}/C^m) in this case. Interestingly, we notice that the relative mismatch costs given by Line (1) and Line (5) are very similar. This suggests that optimizing over the IROC curve alone (and leaving GP and ED fees unchanged) already achieves first-best in most cases.[12]

8. Conclusion and Policy Implications

Acute care systems and patients have long suffered from the mismatch between supply and demand of acute care resources: when GP-type patients seek care at an ED, they use costly medical resources unnecessarily and worsen the ED overcrowding problem; when ED-type patients visit GPs, they need to be referred to an ED, incurring additional costs at the GP and experiencing treatment delay. While acute care systems have implemented incentive mechanisms (e.g., fees, priority queues) so that self-triaged GP-type patients do not go straight to an ED and self-triaged ED-type patients do not head to a GP first, such measures cannot address the fundamental inefficiency resulting from self-triage inaccuracy. Capitalizing on advances in AI technology, virtual triage tools are being developed and deployed worldwide to help improve patient self-triage by providing instantaneous and costless triage recommendations before patients seek care. However, up to this point, the operational implications of adopting virtual triage have not been explored. In this paper, we have sought to fill this gap by studying the impact of virtual triage.

We show that virtual triage has an unintended effect on the care-seeking behavior of *patients*: due to the decentralized nature of virtual triage technology, systems that excessively recommend emergency (primary) care counterintuitively lead to a decrease in ED (GP) visits. Moreover, virtual triage of higher accuracy might increase the equilibrium social cost. Consequently, we show that the adoption of informative virtual triage can lead to a higher equilibrium social cost than before, i.e., in the absence of virtual triage, even when the triage algorithm has reasonably good accuracy. Our analysis therefore suggests that, when evaluating a virtual triage tool, *regulators* should be aware of the potential unintended consequences and carefully assess the impact of the technology on patient care-seeking behavior and equilibrium social cost prior to implementation. Furthermore, as automatically updating a virtual triage algorithm to a more accurate version without re-evaluation and additional regulatory approval may reverse any previously demonstrated benefit, regulators may require a system for certifying these technologies with each new iteration.

To resolve the problems associated with virtual triage and allow healthcare systems to access its operational benefits, we identify two sources of inefficiency in equilibrium outcomes, and propose associated policy actions to alleviate the inefficiency. First, for *healthcare providers*, we find that current GP and ED fees, which induce optimal patient flow in the absence of virtual triage, may fail to induce corresponding optimal patient flow after the adoption of virtual triage. Second, we characterize *technology providers'* optimal decisions on virtual triage accuracy subject to a given IROC curve. We show that the optimal accuracy critically depends on and is nonmonotone in the triage capability of virtual triage. In fact, it may be optimal to have rather large under-triage or over-triage probability unless the triage capability is sufficiently high. Hence, we recommend that

regulators, healthcare providers, and technology companies work closely together to choose the accuracy of the virtual triage algorithm.

In this paper we have abstracted away from the specific parameter values to show, in a general way, how the acute care system's performance may be affected by the introduction of a virtual triage technology. In practice, however, it may be difficult to estimate several of these parameters, making it challenging to optimize over the IROC curve as described in Section 7.2. Instead, virtual triage providers may choose a scoring function following certain heuristics. As noted above, this could actually lead to worse system performance than in the absence of virtual triage. When the triage capability of the virtual triage algorithm is sufficiently high, however, we have seen in Section 7.3 that these potential adverse effects disappear. Hence, our findings suggest that implementing a virtual triage tool may not always be advisable, especially if estimating the system parameters is difficult, or if the triage capability of virtual triage is not particularly high.

In sum, our paper provides a theoretical analysis of the impact of virtual triage and policy actions to optimize its operational benefits. We emphasize decentralized behavior, incentive alignment, and decision on accuracy subject to a given IROC curve to facilitate the adoption of predictive technology to better deliver care to patients. On a broader level, our work highlights the importance of studying nuanced operational details to realize the intended benefits of emerging technologies.

Endnotes

1. Virtual over-triage (under-triage) probability is defined as the probability of a patient requiring primary (emergency) care being recommended to the ED (GP) by virtual triage. Throughout the paper, we omit "virtual" and refer to them as over-triage and under-triage probabilities.

2. Our model can be readily extended to the case where disutilities of waiting per unit time vary across different locations of care, namely, w_G at a GP and w_E at an ED. All of the results will remain unchanged.

3. Note that we do not explicitly include the social cost of non-strategic patients in $C_s(f_{\hat{L}}, f_{\hat{H}})$. This is because the care-seeking behavior and social cost of non-strategic patients are not affected by the introduction of virtual triage or by the behavior of strategic patients. In particular, for the non-strategic patients at the ED, i.e., patients in high-risk situations or requiring immediate life-saving interventions, the disutility of waiting remains the same: upon arrival at the ED, they will be assigned an Emergency Severity Index (ESI) level of 1 or 2, which prioritizes their treatment over that of patients with moderate acute conditions who will be assigned an ESI level of 3, 4, or 5 (Gilboy et al. 2020). Studies have shown patients with moderate acute conditions have clinically negligible effect on the waiting time of patients of ESI level 1 or 2 (Schull et al. 2007, Zane 2007). Meanwhile, a low-acuity patient visiting a GP will rarely be allocated to same-day capacity that is reserved for acute care patients and instead will typically have to schedule their appointment in advance of their visit. Hence, effectively, the social cost of non-strategic patients remains unchanged, regardless of the introduction of virtual triage or the behavior of strategic patients.

4. Equivalently, we assume that the reservation prices of GP and ED are equal to their marginal operating costs. Our analysis can be readily extended to the case in which reservation prices are higher than operating costs.

5. For simplicity, we refer to the nonatomic Nash equilibrium patient flow as equilibrium patient flow for the rest of the paper.

6. We note that Webb and Mills (2019) capture a similar trade-off in a centralized setting; in this paper, we explore the effect of such an accuracy trade-off on patients' care-seeking behavior in a decentralized setting.

7. Note that the original ROC curve can be trivially recovered by vertically reflecting the IROC curve over the line $\alpha = 0.5$.

8. We define self-triage decision as the patient's choice of care in the absence of virtual triage, as characterized in Section 3.

9. While Figure 2 is generated from a specific set of system parameter values, the eight equilibrium regimes and their relative positions are robust to different sets of system parameter values.

10. Parameter values for Figure 3: $b_{\hat{L}} = 0.20, b_{\hat{H}} = 0.75$ under low self-triage accuracy; $b_{\hat{L}} = 0.15, b_{\hat{H}} = 0.80$ under high self-triage accuracy; $\lambda_{\hat{L}} = 30, \lambda_{\hat{H}} = 7$; $a_G = 40, a_E = 100$; $w = 7$; $Q_G = 12, Q_E = 0.05\lambda_E^2$.

11. Parameter values for Figure 4 (right) and Figure 5: $b_{\hat{L}} = 0.20, b_{\hat{H}} = 0.75$; $\lambda_{\hat{L}} = 30, \lambda_{\hat{H}} = 7$; $a_G = 40, a_E = 100$; $w = 7$; $Q_G = 12, Q_E = 0.05\lambda_E^2$.

12. The intuition behind follows from the observation in Figure 4 (right) that by optimizing over IROC curve alone with GP and ED fees unchanged, the equilibrium patient flow tends to be pure strategy, which does not suffer from the potential inefficiency of mixed strategy equilibria. In this case, jointly optimizing over IROC curve and patient flow does not require adjusting the ED fee.

References

Alizamir S, de Véricourt F, Sun P. 2013. Diagnostic accuracy under congestion. *Management Sci.* 59(1):157–171.

Babic B, Gerke S, Evgeniou T, Cohen IG. 2019. Algorithms on regulatory lockdown in medicine. *Science* 366(6470):1202–1204.

Bavafa H, Hitt LM, Terwiesch C. 2018. The impact of e-visits on visit frequencies and patient health: evidence from primary care. *Management Sci.* 64(12):5461–5480.

Bavafa H, Savin S, Terwiesch C. 2019. Redesigning primary care delivery: customized office revisit intervals and e-visits. *Working Paper.*

Blue Shield of California. 2020. COVID-19: virtual triage can help patients, ease burden on hospitals. (April 6). https://www.bcbs.com/the-health-of-america/articles/covid-19-virtual-triage-can-help-patients-ease-burden-hospitals

Boute RN, Van Mieghem JA. 2020. Digital operations: autonomous automation and the smart execution of work. *Working Paper.*

Çakıcı ÖE, Mills AF. 2020. On the role of teletriage in healthcare demand management. *Manufacturing Service Oper. Management*, forthcoming.

Chambers D, Cantrell AJ, Johnson M, Preston L, Baxter SK, Booth A, Turner J. 2019. Digital and online symptom checkers and health assessment/triage services for urgent health problems: systematic review. *BMJ Open* 9:e027743.

Coons KC, DuMoulin JP. 2000. Telephone triage. Technical report, American College of Physicians–American Society of Internal Medicine, Washington, DC.

Corl K. 2019. Hospitals' new emergency department triage systems boost profits but compromise care. *STAT* (September 5). https://www.statnews.com/2019/09/05/triage-system-boost-profits-compromises-care/.

Coyle AL. 2017. Dealing with patient self-triage. *Nursing* 47(12):17–18.

Cui S, Veeraraghavan S. 2016. Blind queues: the impact of consumer beliefs on revenues and congestion. *Management Sci.* 62(12):3656–3672.

Dalton J. 2020. Coronavirus: callers to NHS 111 phone line wait hours and get cut off without being able to speak to nurse. *INDEPENDENT* (March 13). https://www.independent.co.uk/news/uk/home-news/coronavirus-uk-symptoms-nhs-111-phone-line-nurse-a9400351.html.

Debo L, Parlour C, Rajan U. 2012. Signaling quality via queues. *Management Sci.* 58(5):876–891.

Freeman M, Savva N, Scholtes S. 2017. Gatekeepers at work: an empirical analysis of a maternity unit. *Management Sci.* 63(10):3147–3167.

Freeman M, Robinson S, Scholtes S. 2020. Gatekeeping, fast and slow: an empirical study of referral errors in the emergency department. *Management Sci.*, forthcoming.

Gilboy N, Tanabe P, Travers D, Rosenau AM. 2020. Emergency severity index, version 4: implementation handbook. Emergency Nurses Association, Schaumburg, IL.

Gneiting T. 2011. Making and evaluating point forecasts. *Journal of the American Statistical Association* 106(494):746–762.

Grand View Research. 2019. Acute hospital care market growth & trends. https://www.grandviewresearch.com/press-release/global-acute-hospital-care-market

Gupta D, Wang L. 2008. Revenue management for a primary-care clinic in the presence of patient choice. *Oper. Res.* 56(3):576–592.

Hao K. 2020. Doctors are using AI to triage COIVD-19 patients. The tools may be here to stay. *MIT Technology Review* (April 23). https://www.technologyreview.com/2020/04/23/1000410/ai-triage-covid-19-patients-health-care/?truid=cb0787e5baf82e6f6cc19ac58536e5b4&utm_source=the_download&utm_medium=email&utm_campaign=the_download.unpaid.engagement&utm_content=04-24-2020.

Hasija S, Pinker E, Shumsky R. 2005. Staffing and routing in a two-tier call centre. *Internat. J. Oper. Res.* 1(1/2):8–29.

Heaven WD. 2020. Google's medical AI was super accurate in a lab. Real life was a different story. *MIT Technology Review* (April 27). https://www.technologyreview.com/2020/04/27/1000658/google-medical-ai-accurate-lab-real-life-clinic-covid-diabetes-retina-disease/?truid=cb0787e5baf82e6f6cc19ac58536e5b4&utm_source=the_algorithm&utm_medium=email&utm_campaign=the_algorithm.unpaid.engagement&utm_content=05-01-2020

Hirshon JM, Risko N, Calvello EJ, de Ramirez SS, Narayan M, Theodosis C, O'Neill J. 2013. Health systems and services: the role of acute care. *Bulletin of the World Health Organization* 91(5):386–388.

Hu M, Li Y, Wang J. 2018. Efficient ignorance: information heterogeneity in a queue. *Management Sci.* 64(6):2650–2671.

Huang J, Carmeli B, Mandelbaum A. 2012. Control of patient flow in emergency departments, or multiclass queues with deadlines and feedback. *Oper. Res.* 63(4):892–908.

Iserson KV, Moskop JC. 2007. Triage in medicine, part i: concept, history, and types. *Annals of Emergency Medicine* 49(3):275–281.

Kamali MF, Tezcan T, Yildiz O. 2019. When to use provider triage in emergency departments. *Management Sci.* 65(3):1003–1019.

Khalik S. 2014. 2 public hospitals, SGH and KTPH, increase fees for emergency services. *The Straits Times* (April 12). https://www.straitstimes.com/singapore/health/2-public-hospitals-sgh-and-ktph-increase-fees-for-emergency-services.

Kocher KE, Ayanian JZ. 2016. A fractured system: where do you go when you suddenly need health care? *The Conversation* (October 31). https://theconversation.com/a-fractured-system-where-do-you-go-when-you-suddenly-need-health-care-66662.

Lee H, Pinker E, Shumsky R. 2012. Outsourcing a two-level service process. *Management Sci.* 58(8):1569–1584.

Lega F, Mengoni A. 2008. Why non-urgent patients choose emergency over primary care services? Empirical evidence and managerial implications. *Health Policy* 88(2):326–338.

Levi R, Magnanti T, Shaposhnik Y. 2019. Scheduling with testing. *Management Sci.* 65(2):776–793.

Liu Y, Wang X, Gilbert S, Lai G. 2018. Pricing, quality and competition at on-demand healthcare service platforms. *Working Paper.*

Lovett L. 2018. AI triage chatbots trekking toward a standard of care despite criticism. *Mobile Health News* (November 2). https://www.mobihealthnews.com/content/ai-triage-chatbots-trekking-toward-standard-care-despite-criticism.

Meyer A, Giardina TD, Spitzmueller C, Shahid U, Scott T, Singh H. 2020. Patient perspectives on the usefulness of an artificial intelligence–assisted symptom checker: cross-sectional survey study. *J Med Internet Res* 22(1):e14679.

Papanastasiou Y, Bakshi N, Savva N. 2015. Scarcity strategies under quasi-Bayesian social learning. *Working Paper.*

Papanastasiou Y, Bimpikis K, Savva N. 2018. Crowdsourcing exploration. *Management Sci.* 64(4):1727–1746.

Rajan B, Tezcan T, Seidmann A. 2019. Service systems with heterogeneous customers: investigating the effect of telemedicine on chronic care. *Management Sci.* 65(3):1236–1267.

Saghafian S, Hopp WJ, Van Oyen MP, Desmond JS, Kronick SL. 2012. Patient streaming as a mechanism for improving responsiveness in emergency departments. *Oper. Res.* 60(5):1080–1097.

Savin S, Xu Y, Zhu L. 2019. Delivering multi-specialty care via online telemedicine platforms. *Working Paper.*

Schmeidler D. 1973. Equilibrium points of nonatomic games. *J. Statist. Phys.* 7(4):295–300.

Schull MJ, Kiss A, Szalai J. 2007. The effect of low-complexity patients on emergency department waiting times. *Annals of Emergency Medicine* 49(3):257–264.

Semigran HL, Linder JA, Gidengil C, Mehrotra A. 2015. Evaluation of symptom checkers for self diagnosis and triage: audit study. *BMJ* 351:h3480.

Sharma S, Xu Y, Gupta MK, Courcoubetis C. 2019. Non-urgent visits and emergency department congestion: patients' choice and incentive mechanisms. *Working Paper.*

Shumsky R, Pinker E. 2003. Gatekeepers and referrals in services. *Management Sci.* 49(7):839–856.

Singh S, Gurvich I, Van Mieghem JA. 2020. Feature-based design of priority queues: digital triage in healthcare. *Working Paper.*

Sun Z, Argon NT, Ziya S. 2018. Patient triage and prioritization under austere conditions. *Management Sci.* 64(10):4471–4489.

Tencent. 2020. LancTencent open sources COVID-19 self-triage assistant. https://www.tencent.com/en-us/responsibility/combat-covid-19-assistant-module.html.

Trivedi S, Littmann J, Kapur P, Betz M, Stempien J. 2017. LO09: assessing the ability of emergency department patients to self-triage by using an electronic questionnaire: a pilot study. *CJEM* 19(S1):S30–S30.

Veeraraghavan S, Debo L. 2009. Joining longer queues: information externalities in queue choice. *Manufacturing Service Oper. Management* 11(4):543–562.

Verzantvoort NCM, Teunis T, Verheij TJM, van der Velden A. 2018. Self-triage for acute primary care via a smartphone application: practical, safe and efficient? *PLOS ONE* 13(6):e0199284.

Webb EM, Mills AF. 2019. Incentive-compatible prehospital triage in emergency medical services. *Prod. Oper. Manag.* 28(9):2221–2241.

Winn AN, Somai M, Fergestrom N, Crotty BH. 2019. Association of use of online symptom checkers with patients' plans for seeking care. *JAMA Network Open* 2(12):e1918561.

Zane RD. 2007. Are low-acuity patients clogging up the ED? *NEJM Journal Watch* Reviewing Schull et al. 2007 Ann Emerg Med 2007 Mar.

Zayas-Cabán G, Xie J, Green LV, Lewis ME. 2014. Optimal control of an emergency room triage and treatment process. *Working Paper.*

Zorc S, Chick SE, Hasija S. 2017. Outcomes-based reimbursement policies for chronic care pathways. *Working Paper.*

Online Technical Appendix

In this e-companion we provide technical results that are required for our analysis, as well as detailed proofs of all the mathematical results in the paper.

EC.1. Characterization of Optimal and Equilibrium Patient Flow

Suppose that there are n types of patients seeking acute care. Patients of type i have an arrival rate λ_i, with a probability b_i of being ED-type. WLOG, we assume $b_1 < b_2 < ... < b_n$. Suppose the expected GP fee per visit is p_G, and the expected ED fee per visit is p_E. Let $f_i \in [0,1]$ denote the probability of type i patients visiting the ED directly.

EC.1.1. Optimal Patient Flow

Let $C_s(\boldsymbol{f})$ denote the social cost, where $\boldsymbol{f} = (f_1, f_2, ..., f_n)$, and it can be expressed as follows:

$$C_s(\boldsymbol{f}) = \lambda_G(\boldsymbol{f}) w Q_G + \lambda_E(\boldsymbol{f}) w Q_E(\lambda_E(\boldsymbol{f})) + S_G(\lambda_G(\boldsymbol{f})) + S_E(\lambda_E(\boldsymbol{f})) \tag{EC.1}$$

where patient arrival rate to GP $\lambda_G(\boldsymbol{f}) = \sum_{i=1}^{n}(1-f_i)\lambda_i$, and patient arrival rate to ED $\lambda_E(\boldsymbol{f}) = \sum_{i=1}^{n}(b_i(1-f_i)\lambda_i + f_i\lambda_i)$.

LEMMA EC.1. *$C_s(\boldsymbol{f})$ is jointly convex in $f_1, f_2, ..., f_n$.*

Proof of Lemma EC.1 We first have $\lambda_G(\boldsymbol{f}) = \sum_{i=1}^{n}(1-f_i)\lambda_i$ and $\lambda_E(\boldsymbol{f}) = \sum_{i=1}^{n}(b_i(1-f_i)\lambda_i + f_i\lambda_i)$ being linear functions in $f_1, f_2, ..., f_n$. Now we consider the term $\lambda_E(\boldsymbol{f}) w Q_E(\lambda_E(\boldsymbol{f}))$ in Equation EC.1. The first-order partial derivative w.r.t $\lambda_E(\boldsymbol{f})$ is

$$\frac{\partial \lambda_E(\boldsymbol{f}) w Q_E(\lambda_E(\boldsymbol{f}))}{\partial \lambda_E(\boldsymbol{f})} = w Q_E(\lambda_E(\boldsymbol{f})) + \lambda_E(\boldsymbol{f}) w \frac{\partial Q_E(\lambda_E(\boldsymbol{f}))}{\partial \lambda_E(\boldsymbol{f})} > 0$$

while the second-order partial derivative is

$$\frac{\partial^2 \lambda_E(\boldsymbol{f}) w Q_E(\lambda_E(\boldsymbol{f}))}{\partial \lambda_E(\boldsymbol{f})^2} = 2w \frac{\partial Q_E(\lambda_E(\boldsymbol{f}))}{\partial \lambda_E(\boldsymbol{f})} + \lambda_E(\boldsymbol{f}) w \frac{\partial^2 Q_E(\lambda_E(\boldsymbol{f}))}{\partial \lambda_E(\boldsymbol{f})^2} > 0$$

Hence, $\lambda_E(\boldsymbol{f}) w Q_E(\lambda_E(\boldsymbol{f}))$ is convex in $\lambda_E(\boldsymbol{f})$. As $\lambda_E(\boldsymbol{f})$ is linear in $f_1, f_2, ..., f_n$, by preservation of convexity, $\lambda_E(\boldsymbol{f}) w Q_E(\lambda_E(\boldsymbol{f}))$ is jointly convex in $f_1, f_2, ..., f_n$. As the remaining terms in Equation EC.1 are linear in $f_1, f_2, ..., f_n$, $C_s(\boldsymbol{f})$ is jointly convex in $f_1, f_2, ..., f_n$. □

Let $\boldsymbol{f^*} = (f_1^*, f_2^*, ..., f_n^*)$ denote the unique solution to the following problem:

$$\min_{0 \le f_1, f_2, ..., f_n \le 1} C_s(\boldsymbol{f}) \tag{EC.2}$$

LEMMA EC.2. *$\boldsymbol{f^*}$ satisfies the following structural property: $\exists\ i \in \{1, 2, ..., n\}$ s.t. $f_i^* \in [0,1], f_j^* = 0, \forall\ j < i$, and $f_k^* = 1, \forall\ k > i$.*

Proof of Lemma EC.2 Case 1: Suppose $\exists\ i \in \{1,2,...,n\}$ s.t. $f_i^* \in (0,1)$. Then we have

$$\frac{\partial C_s(\boldsymbol{f})}{\partial f_i}\bigg|_{\boldsymbol{f}^*} = -(a_G + wQ_G)\lambda_i + a_E(1-b_i)\lambda_i + (1-b_i)\lambda_i wQ_E(\lambda_E(\boldsymbol{f}^*)) + \lambda_E(\boldsymbol{f}^*)w(1-b_i)\lambda_i \frac{\partial Q_E(\lambda_E(\boldsymbol{f}))}{\partial \lambda_E(\boldsymbol{f})}\bigg|_{\boldsymbol{f}^*} = 0$$

We have $b_j < b_i, \forall\ j < i$, and therefore

$$\frac{\partial C_s(\boldsymbol{f})}{\partial f_j}\bigg|_{\boldsymbol{f}^*} = -(a_G + wQ_G)\lambda_j + a_E(1-b_j)\lambda_j + (1-b_j)\lambda_j wQ_E(\lambda_E(\boldsymbol{f}^*)) + \lambda_E(\boldsymbol{f}^*)w(1-b_j)\lambda_j \frac{\partial Q_E(\lambda_E(\boldsymbol{f}))}{\partial \lambda_E(\boldsymbol{f})}\bigg|_{\boldsymbol{f}^*} > 0$$

Hence, we have $f_j^* = 0, \forall\ j < i$. Similarly, we have $b_k > b_i, \forall\ k > i$, and therefore

$$\frac{\partial C_s(\boldsymbol{f})}{\partial f_k}\bigg|_{\boldsymbol{f}^*} = -(a_G + wQ_G)\lambda_k + a_E(1-b_k)\lambda_k + (1-b_k)\lambda_k wQ_E(\lambda_E(\boldsymbol{f}^*)) + \lambda_E(\boldsymbol{f}^*)w(1-b_k)\lambda_k \frac{\partial Q_E(\lambda_E(\boldsymbol{f}))}{\partial \lambda_E(\boldsymbol{f})}\bigg|_{\boldsymbol{f}^*} < 0$$

Hence, we have $f_k^* = 1, \forall\ k > i$.

Case 2: Suppose $\nexists\ i \in \{1,2,...,n\}$ s.t. $f_i^* \in (0,1)$. $\forall\ i$ s.t. $f_i^* = 0$, we have

$$\frac{\partial C_s(\boldsymbol{f})}{\partial f_i}\bigg|_{\boldsymbol{f}^*} = -(a_G + wQ_G)\lambda_i + a_E(1-b_i)\lambda_i + (1-b_i)\lambda_i wQ_E(\lambda_E(\boldsymbol{f}^*)) + \lambda_E(\boldsymbol{f}^*)w(1-b_i)\lambda_i \frac{\partial Q_E(\lambda_E(\boldsymbol{f}))}{\partial \lambda_E(\boldsymbol{f})}\bigg|_{\boldsymbol{f}^*} \geq 0$$

We have $b_j < b_i, \forall\ j < i$, and therefore

$$\frac{\partial C_s(\boldsymbol{f})}{\partial f_j}\bigg|_{\boldsymbol{f}^*} = -(a_G + wQ_G)\lambda_j + a_E(1-b_j)\lambda_j + (1-b_j)\lambda_j wQ_E(\lambda_E(\boldsymbol{f}^*)) + \lambda_E(\boldsymbol{f}^*)w(1-b_j)\lambda_j \frac{\partial Q_E(\lambda_E(\boldsymbol{f}))}{\partial \lambda_E(\boldsymbol{f})}\bigg|_{\boldsymbol{f}^*} > 0$$

Hence, we have $f_j^* = 0, \forall\ j < i$. Similarly, $\forall\ i$ s.t. $f_i^* = 1$, we have

$$\frac{\partial C_s(\boldsymbol{f})}{\partial f_i}\bigg|_{\boldsymbol{f}^*} = -(a_G + wQ_G)\lambda_i + a_E(1-b_i)\lambda_i + (1-b_i)\lambda_i wQ_E(\lambda_E(\boldsymbol{f}^*)) + \lambda_E(\boldsymbol{f}^*)w(1-b_i)\lambda_i \frac{\partial Q_E(\lambda_E(\boldsymbol{f}))}{\partial \lambda_E(\boldsymbol{f})}\bigg|_{\boldsymbol{f}^*} \leq 0$$

We have $b_k > b_i, \forall\ k > i$, and therefore

$$\frac{\partial C_s(\boldsymbol{f})}{\partial f_k}\bigg|_{\boldsymbol{f}^*} = -(a_G + wQ_G)\lambda_k + a_E(1-b_k)\lambda_k + (1-b_k)\lambda_k wQ_E(\lambda_E(\boldsymbol{f}^*)) + \lambda_E(\boldsymbol{f}^*)w(1-b_k)\lambda_k \frac{\partial Q_E(\lambda_E(\boldsymbol{f}))}{\partial \lambda_E(\boldsymbol{f})}\bigg|_{\boldsymbol{f}^*} < 0$$

Hence, we have $f_k^* = 1, \forall\ k > i$. $\square$

EC.1.2. Equilibrium Patient Flow

LEMMA EC.3. *$\forall\ p_G, p_E \geq 0$, there exists a unique patient flow in equilibrium.*

Proof of Lemma EC.3 We define the following *potential function* (Roughgarden 2007), $\Phi(\boldsymbol{f})$, for our nonatomic game:

$$\Phi(\boldsymbol{f}) = \int_0^{\lambda_G(\boldsymbol{f})} wQ_G dx + \int_0^{\lambda_E(\boldsymbol{f})} wQ_E(x)dx + \lambda_G(\boldsymbol{f})p_G + \lambda_E(\boldsymbol{f})p_E \tag{EC.3}$$

Let $\boldsymbol{f^e} = (f_1^e, f_2^e, ..., f_n^e)$ denote equilibrium patient flow, which is the solution to the following problem:

$$\min_{0\leq f_1, f_2, ..., f_n \leq 1} \Phi(\boldsymbol{f}) \tag{EC.4}$$

The first-order partial derivative of $\int_0^{\lambda_E(\boldsymbol{f})} wQ_E(x)dx$ w.r.t $\lambda_E(\boldsymbol{f})$ is

$$\frac{\partial(\int_0^{\lambda_E(\boldsymbol{f})} wQ_E(x)dx)}{\partial \lambda_E(\boldsymbol{f})} = wQ_E(\lambda_E(\boldsymbol{f})) > 0$$

while the second-order partial derivative is

$$\frac{\partial^2(\int_0^{\lambda_E(\boldsymbol{f})} wQ_E(x)dx)}{\partial \lambda_E(\boldsymbol{f})^2} = w\frac{\partial Q_E(\lambda_E(\boldsymbol{f}))}{\partial \lambda_E(\boldsymbol{f})} > 0$$

Hence $\int_0^{\lambda_E(\boldsymbol{f})} wQ_E(x)dx$ is convex in $\lambda_E(\boldsymbol{f})$. As $\lambda_E(\boldsymbol{f})$ is linear in $f_1, f_2, ..., f_n$, by preservation of convexity, $\int_0^{\lambda_E(\boldsymbol{f})} wQ_E(x)dx$ is jointly convex in $f_1, f_2, ..., f_n$. As the remaining terms in Equation EC.3 are linear in $f_1, f_2, ..., f_n$, $\Phi(\boldsymbol{f})$ is jointly convex in $f_1, f_2, ..., f_n$. Hence, there is a unique solution to the problem EC.4. □

LEMMA EC.4. *$\boldsymbol{f^e}$ satisfies the following structural property: $\exists\ i \in \{1, 2, ..., n\}$ s.t. $f_i^e \in [0, 1], f_j^e = 0, \forall\ j < i$, and $f_k^e = 1, \forall\ k > i$.*

Proof of Lemma EC.4 Case 1: Suppose $\exists\ i \in \{1, 2, ..., n\}$ s.t. $f_i^e \in (0, 1)$. Then we have

$$\left.\frac{\partial \Phi(\boldsymbol{f})}{\partial f_i}\right|_{\boldsymbol{f^e}} = -(p_G + wQ_G)\lambda_i + (1 - b_i)\lambda_i(p_E + wQ_E(\lambda_E(\boldsymbol{f^e}))) = 0$$

We have $b_j < b_i, \forall\ j < i$, and therefore

$$\left.\frac{\partial \Phi(\boldsymbol{f})}{\partial f_j}\right|_{\boldsymbol{f^e}} = -(p_G + wQ_G)\lambda_j + (1 - b_j)\lambda_j(p_E + wQ_E(\lambda_E(\boldsymbol{f^e}))) > 0$$

Hence, we have $f_j^e = 0, \forall\ j < i$. Similarly, we have $b_k > b_i, \forall\ k > i$, and therefore

$$\left.\frac{\partial \Phi(\boldsymbol{f})}{\partial f_k}\right|_{\boldsymbol{f^e}} = -(p_G + wQ_G)\lambda_k + (1 - b_k)\lambda_k(p_E + wQ_E(\lambda_E(\boldsymbol{f^e}))) < 0$$

Hence, we have $f_k^e = 1, \forall\ k > i$.

Case 2: Suppose $\nexists\ i \in \{1,2,...,n\}$ s.t. $f_i^e \in (0,1)$. $\forall\ i$ s.t. $f_i^e = 0$, we have

$$\left.\frac{\partial \Phi(\boldsymbol{f})}{\partial f_i}\right|_{\boldsymbol{f^e}} = -(p_G + wQ_G)\lambda_i + (1-b_i)\lambda_i(p_E + wQ_E(\lambda_E(\boldsymbol{f^e}))) \geq 0$$

We have $b_j < b_i, \forall\ j < i$, and therefore

$$\left.\frac{\partial \Phi(\boldsymbol{f})}{\partial f_j}\right|_{\boldsymbol{f^e}} = -(p_G + wQ_G)\lambda_j + (1-b_j)\lambda_j(p_E + wQ_E(\lambda_E(\boldsymbol{f^e}))) > 0$$

Hence, we have $f_j^e = 0, \forall\ j < i$. Similarly, $\forall\ i$ s.t. $f_i^e = 1$, we have

$$\left.\frac{\partial \Phi(\boldsymbol{f})}{\partial f_i}\right|_{\boldsymbol{f^e}} = -(p_G + wQ_G)\lambda_i + (1-b_i)\lambda_i(p_E + wQ_E(\lambda_E(\boldsymbol{f^e}))) \leq 0$$

We have $b_k > b_i, \forall\ k > i$, and therefore

$$\left.\frac{\partial \Phi(\boldsymbol{f})}{\partial f_k}\right|_{\boldsymbol{f^e}} = -(p_G + wQ_G)\lambda_k + (1-b_k)\lambda_k(p_E + wQ_E(\lambda_E(\boldsymbol{f^e}))) < 0$$

Hence, we have $f_k^e = 1, \forall\ k > i$. □

LEMMA EC.5. *Let p_{G1} be the expected GP fee and p_{E1} be the associated minimum expected ED fee that induce a specific equilibrium patient flow $\boldsymbol{f^e}$. Let p_{G2} be another expected GP fee and p_{E2} be the associated minimum expected ED fee that induce the same equilibrium patient flow $\boldsymbol{f^e}$. If $p_{G2} > p_{G1}$, we have $p_{E2} > p_{E1}$.*

Proof of Lemma EC.5 Case 1: Suppose $\exists\ i \in \{1,2,...,n\}$ s.t. $f_i^e \in (0,1)$. Then for a given p_{G1}, the associated p_{E1} is given by the solution of

$$\left.\frac{\partial \Phi(\boldsymbol{f})}{\partial f_i}\right|_{\boldsymbol{f^e}} = -(p_{G1} + wQ_G)\lambda_i + (1-b_i)\lambda_i(p_{E1} + wQ_E(\lambda_E(\boldsymbol{f^e}))) = 0$$

Similarly, for a given p_{G2}, the associated p_{E2} is given by the solution of

$$\left.\frac{\partial \Phi(\boldsymbol{f})}{\partial f_i}\right|_{\boldsymbol{f^e}} = -(p_{G2} + wQ_G)\lambda_i + (1-b_i)\lambda_i(p_{E2} + wQ_E(\lambda_E(\boldsymbol{f^e}))) = 0$$

It is clear that if $p_{G2} > p_{G1}$, we have $p_{E2} > p_{E1}$.

Case 2: Suppose $\nexists\ i \in \{1,2,...,n\}$ s.t. $f_i^e \in (0,1)$. Suppose $f_j^e = 0$ and $f_{j+1}^e = 1$. We then have

$$\left.\frac{\partial \Phi(\boldsymbol{f})}{\partial f_j}\right|_{\boldsymbol{f^e}} = -(p_G + wQ_G)\lambda_j + (1-b_j)\lambda_j(p_E + wQ_E(\lambda_E(\boldsymbol{f^e}))) \geq 0$$

$$\left.\frac{\partial \Phi(\boldsymbol{f})}{\partial f_{j+1}}\right|_{\boldsymbol{f^e}} = -(p_G + wQ_G)\lambda_{j+1} + (1-b_{j+1})\lambda_{j+1}(p_E + wQ_E(\lambda_E(\boldsymbol{f^e}))) \leq 0$$

Then for a given p_{G1}, the associated p_{E1} is given by the solution of

$$\left.\frac{\partial \Phi(\boldsymbol{f})}{\partial f_j}\right|_{\boldsymbol{f^e}} = -(p_{G1} + wQ_G)\lambda_j + (1-b_j)\lambda_j(p_{E1} + wQ_E(\lambda_E(\boldsymbol{f^e}))) = 0$$

Similarly, for a given p_{G2}, the associated p_{E2} is given by the solution of

$$\left.\frac{\partial \Phi(\boldsymbol{f})}{\partial f_j}\right|_{\boldsymbol{f^e}} = -(p_{G2} + wQ_G)\lambda_j + (1-b_j)\lambda_j(p_{E2} + wQ_E(\lambda_E(\boldsymbol{f^e}))) = 0$$

It is clear that if $p_{G2} > p_{G1}$, we have $p_{E2} > p_{E1}$. □

EC.2. Proofs

This section provides detailed proofs of all the mathematical results in the paper.

EC.2.1. Proofs for Section 3

Proof of Proposition 1 $C_s(f_{\hat{L}}, f_{\hat{H}})$ is jointly convex in $f_{\hat{L}}$ and $f_{\hat{H}}$ by Lemma EC.1 with $n=2$. In addition, by Lemma EC.2, the unique $f_{\hat{L}}$ and $f_{\hat{H}}$ that minimize $C_s(f_{\hat{L}}, f_{\hat{H}})$, $f^*_{\hat{L}}$ and $f^*_{\hat{H}}$, take one of the following forms: (1) $f^*_{\hat{L}}=0, f^*_{\hat{H}}=0$; (2) $f^*_{\hat{L}}=0, f^*_{\hat{H}}\in(0,1)$; (3) $f^*_{\hat{L}}=0, f^*_{\hat{H}}=1$; (4) $f^*_{\hat{L}}\in(0,1), f^*_{\hat{H}}=1$; (5) $f^*_{\hat{L}}=1, f^*_{\hat{H}}=1$. The necessary and sufficient conditions for $f^*_{\hat{L}}=0$ and $f^*_{\hat{H}}=1$ are

$$\begin{aligned}\left.\frac{\partial C_s(f_{\hat{L}}, f_{\hat{H}})}{\partial f_{\hat{L}}}\right|_{f_{\hat{L}}=0, f_{\hat{H}}=1} &= -(a_G+wQ_G)\lambda_{\hat{L}} + (1-b_{\hat{L}})\lambda_{\hat{L}} M_E(b_{\hat{L}}\lambda_{\hat{L}}+\lambda_{\hat{H}}) \geq 0\\ &\Leftrightarrow (1-b_{\hat{L}})M_E(b_{\hat{L}}\lambda_{\hat{L}}+\lambda_{\hat{H}}) \geq a_G+wQ_G\\ &\Leftarrow (1-b_{\hat{L}})M_E(\lambda_H) \geq a_G+wQ_G\end{aligned}$$

which is implied by Assumption 1 (i), and

$$\begin{aligned}\left.\frac{\partial C_s(f_{\hat{L}}, f_{\hat{H}})}{\partial f_{\hat{H}}}\right|_{f_{\hat{L}}=0, f_{\hat{H}}=1} &= -(a_G+wQ_G)\lambda_{\hat{H}} + (1-b_{\hat{H}})\lambda_{\hat{H}} M_E(b_{\hat{L}}\lambda_{\hat{L}}+\lambda_{\hat{H}}) \leq 0\\ &\Leftrightarrow (1-b_{\hat{H}})M_E(b_{\hat{L}}\lambda_{\hat{L}}+\lambda_{\hat{H}}) \leq a_G+wQ_G\end{aligned}$$

which is implied by Assumption 1 (ii). Hence, the unique minimum social cost is given by $f^*_{\hat{L}}=0$ and $f^*_{\hat{H}}=1$. □

Proof of Proposition 2 The existence and uniqueness of equilibrium patient flow under any $p_G, p_E \geq 0$ directly follow from Lemma EC.3 with $n=2$. We now show that $\hat{p}^*_G$ and $\hat{p}^*_E$ align the equilibrium patient flow with the optimal patient flow, i.e., $f^e_{\hat{L}}=f^*_{\hat{L}}=0$ and $f^e_{\hat{H}}=f^*_{\hat{H}}=1$.

Case 1: Suppose we have $\lambda_H w \frac{\partial Q_E(\lambda_E)}{\partial \lambda_E}\Big|_{\lambda_H} + wQ_E(\lambda_H) - wQ_E(b_{\hat{L}}\lambda_{\hat{L}}+\lambda_{\hat{H}}) \leq 0$. Then under $\hat{p}^*_G = a_G$ and $\hat{p}^*_E = a_E$, we have

$$\begin{aligned}\left.\frac{\partial \Phi(f_{\hat{L}}, f_{\hat{H}})}{\partial f_{\hat{L}}}\right|_{f_{\hat{L}}=0, f_{\hat{H}}=1} &= -(\hat{p}^*_G+wQ_G)\lambda_{\hat{L}} + (1-b_{\hat{L}})\lambda_{\hat{L}}[\hat{p}^*_E + wQ_E(b_{\hat{L}}\lambda_{\hat{L}}+\lambda_{\hat{H}})] \geq 0\\ &\Leftrightarrow (1-b_{\hat{L}})[a_E + wQ_E(b_{\hat{L}}\lambda_{\hat{L}}+\lambda_{\hat{H}})] \geq a_G+wQ_G\\ &\Leftarrow (1-b_{\hat{L}})[a_E + \lambda_H w\left.\frac{\partial Q_E(\lambda_E)}{\partial \lambda_E}\right|_{\lambda_H} + wQ_E(\lambda_H)] \geq a_G+wQ_G\\ &\Leftrightarrow (1-b_{\hat{L}})M_E(\lambda_H) \geq a_G+wQ_G\end{aligned}$$

which is implied by Assumption 1 (i), and

$$\begin{aligned}\left.\frac{\partial \Phi(f_{\hat{L}}, f_{\hat{H}})}{\partial f_{\hat{H}}}\right|_{f_{\hat{L}}=0, f_{\hat{H}}=1} &= -(\hat{p}^*_G+wQ_G)\lambda_{\hat{H}} + (1-b_{\hat{H}})\lambda_{\hat{H}}[\hat{p}^*_E + wQ_E(b_{\hat{L}}\lambda_{\hat{L}}+\lambda_{\hat{H}})] \leq 0\\ &\Leftrightarrow (1-b_{\hat{H}})[a_E + wQ_E(b_{\hat{L}}\lambda_{\hat{L}}+\lambda_{\hat{H}})] \leq a_G+wQ_G\\ &\Leftarrow (1-b_{\hat{H}})M_E(b_{\hat{L}}\lambda_{\hat{L}}+\lambda_{\hat{H}}) \leq a_G+wQ_G\end{aligned}$$

which is implied by Assumption 1 (ii). Hence, $\hat{p}_G^* = a_G$ and $\hat{p}_E^* = a_E$ are the minimum expected GP and ED fees that can recover GP/ED operating costs and induce optimal patient flow with $f_{\hat{L}}^e = f_{\hat{L}}^* = 0$ and $f_{\hat{H}}^e = f_{\hat{H}}^* = 1$.

Case 2: Suppose we have $\lambda_H w \frac{\partial Q_E(\lambda_E)}{\partial \lambda_E}\Big|_{\lambda_H} + wQ_E(\lambda_H) - wQ_E(b_{\hat{L}}\lambda_{\hat{L}} + \lambda_{\hat{H}}) > 0$. Then under $\hat{p}_G^* = a_G$ and $\hat{p}_E^* = a_E + \lambda_H w \frac{\partial Q_E(\lambda_E)}{\partial \lambda_E}\Big|_{\lambda_H} + wQ_E(\lambda_H) - wQ_E(b_{\hat{L}}\lambda_{\hat{L}} + \lambda_{\hat{H}})$, we have

$$\begin{aligned}
\frac{\partial \Phi(f_{\hat{L}}, f_{\hat{H}})}{\partial f_{\hat{L}}}\Big|_{f_{\hat{L}}=0, f_{\hat{H}}=1} &= -(\hat{p}_G^* + wQ_G)\lambda_{\hat{L}} + (1-b_{\hat{L}})\lambda_{\hat{L}}[\hat{p}_E^* + wQ_E(b_{\hat{L}}\lambda_{\hat{L}} + \lambda_{\hat{H}})] \geq 0 \\
&\Leftrightarrow (1-b_{\hat{L}})[a_E + \lambda_H w \frac{\partial Q_E(\lambda_E)}{\partial \lambda_E}\Big|_{\lambda_H} + wQ_E(\lambda_H)] \geq a_G + wQ_G \\
&\Leftrightarrow (1-b_{\hat{L}})M_E(\lambda_H) \geq a_G + wQ_G
\end{aligned}$$

which is implied by Assumption 1 (i), and

$$\begin{aligned}
\frac{\partial \Phi(f_{\hat{L}}, f_{\hat{H}})}{\partial f_{\hat{H}}}\Big|_{f_{\hat{L}}=0, f_{\hat{H}}=1} &= -(\hat{p}_G^* + wQ_G)\lambda_{\hat{H}} + (1-b_{\hat{H}})\lambda_{\hat{H}}[\hat{p}_E^* + wQ_E(b_{\hat{L}}\lambda_{\hat{L}} + \lambda_{\hat{H}})] \leq 0 \\
&\Leftrightarrow (1-b_{\hat{H}})[a_E + \lambda_H w \frac{\partial Q_E(\lambda_E)}{\partial \lambda_E}\Big|_{\lambda_H} + wQ_E(\lambda_H)] \leq a_G + wQ_G \\
&\Leftarrow (1-b_{\hat{H}})M_E(b_{\hat{L}}\lambda_{\hat{L}} + \lambda_{\hat{H}}) \leq a_G + wQ_G
\end{aligned}$$

which is implied by Assumption 1 (ii). In addition, under $\hat{p}_G^* = a_G$, an expected ED fee lower than $\hat{p}_E^* = a_E + \lambda_H w \frac{\partial Q_E(\lambda_E)}{\partial \lambda_E}\Big|_{\lambda_H} + wQ_E(\lambda_H) - wQ_E(b_{\hat{L}}\lambda_{\hat{L}} + \lambda_{\hat{H}})$ could violate Assumption 1 (i). Hence, $\hat{p}_E^* = a_E + \lambda_H w \frac{\partial Q_E(\lambda_E)}{\partial \lambda_E}\Big|_{\lambda_H} + wQ_E(\lambda_H) - wQ_E(b_{\hat{L}}\lambda_{\hat{L}} + \lambda_{\hat{H}})$ is the associated minimum expected ED fee for $\hat{p}_G^* = a_G$ that induces optimal patient flow. Then, by Lemma EC.5, $\hat{p}_G^* = a_G$ and $\hat{p}_E^* = a_E + \lambda_H w \frac{\partial Q_E(\lambda_E)}{\partial \lambda_E}\Big|_{\lambda_H} + wQ_E(\lambda_H) - wQ_E(b_{\hat{L}}\lambda_{\hat{L}} + \lambda_{\hat{H}})$ are the minimum expected GP and ED fees that can recover GP/ED operating costs and induce optimal patient flow with $f_{\hat{L}}^e = f_{\hat{L}}^* = 0$ and $f_{\hat{H}}^e = f_{\hat{H}}^* = 1$. □

EC.2.2. Proofs for Section 4

Proof of Lemma 1. Let $p_H = \int_0^1 sg(s)ds$ denote the fraction of H patients in the patient base and $p_L = 1 - p_H$ denote the fraction of L patients. We have the under-triage probability

$$\alpha(\bar{s}) = Prob(\tilde{L}|H) = \frac{\int_0^{\bar{s}} sg(s)ds}{p_H} \tag{EC.5}$$

and the over-triage probability

$$\beta(\bar{s}) = Prob(\tilde{H}|L) = \frac{\int_{\bar{s}}^1 (1-s)g(s)ds}{p_L} = 1 - \frac{\int_0^{\bar{s}} (1-s)g(s)ds}{p_L} \tag{EC.6}$$

Clearly when $\bar{s}=0$, we have $\alpha(\bar{s})=0, \beta(\bar{s})=1$; when $\bar{s}=1$, we have $\alpha(\bar{s})=1, \beta(\bar{s})=0$. In addition, we have

$$\frac{\partial \alpha}{\partial \beta}=\frac{\partial \alpha}{\partial \bar{s}}\frac{\partial \bar{s}}{\partial \beta}=\frac{\bar{s}g(\bar{s})}{p_H}(-\frac{p_L}{(1-\bar{s})g(\bar{s})})=-\frac{p_L}{p_H}\frac{\bar{s}}{(1-\bar{s})}\leq 0$$

and

$$\frac{\partial^2 \alpha}{\partial \beta^2}=-\frac{p_L}{p_H}\frac{\partial[\bar{s}/(1-\bar{s})]}{\partial \beta}=-\frac{p_L}{p_H}\frac{\partial[\bar{s}/(1-\bar{s})]}{\partial \bar{s}}\frac{\partial \bar{s}}{\partial \beta}=\frac{p_L^2}{p_H(1-\bar{s})^3 g(\bar{s})}\geq 0$$

Hence, $\alpha=r(\beta)$ is a decreasing and convex function in β, with $r(0)=1, r(1)=0$. □

Proof of Corollary 1. $\forall\ b\in[0,1]$, we have

$$b_{\tilde{L}}=\frac{\alpha b}{\alpha b+(1-\beta)(1-b)} \tag{EC.7a}$$

$$b_{\tilde{H}}=\frac{(1-\alpha)b}{(1-\alpha)b+\beta(1-b)} \tag{EC.7b}$$

and therefore

$$b_{\tilde{L}}-b=\frac{b(1-b)(\alpha+\beta-1)}{\alpha b+(1-\beta)(1-b)}\leq 0 \tag{EC.8a}$$

$$b_{\tilde{H}}-b=\frac{(1-\alpha-\beta)b(1-b)}{(1-\alpha)b+\beta(1-b)}\geq 0 \tag{EC.8b}$$

as by Lemma 1, we have $\alpha+\beta\leq 1$.

□

Proof of Lemma 2. $\forall\ \bar{s}_1,\bar{s}_2\in[0,1]$ s.t.

$$\frac{\int_{\bar{s}_1}^1(1-s)g_1(s)ds}{p_L}=\frac{\int_{\bar{s}_2}^1(1-s)g_2(s)ds}{p_L}$$

we have

$$\frac{\int_0^{\bar{s}_1}sg_1(s)ds}{p_H}\geq\frac{\int_0^{\bar{s}_2}sg_2(s)ds}{p_H}$$

Hence, $\forall\ \bar{s}_1,\bar{s}_2\in[0,1]$ s.t. $\beta(\bar{s}_1)=\beta(\bar{s}_2)$, we have $\alpha(\bar{s}_1)\geq\alpha(\bar{s}_2)$. This implies $r_1(\beta)\geq r_2(\beta)$, $\forall\ \beta\in[0,1]$. □

EC.2.3. Proofs for Section 5

Proof of Lemma 3. (i) $\frac{\partial b_{\hat{T}\tilde{L}}}{\partial\alpha}=\frac{b_{\hat{T}}(1-\beta)(1-b_{\hat{T}})}{[\alpha b_{\hat{T}}+(1-\beta)(1-b_{\hat{T}})]^2}>0, \frac{\partial b_{\hat{T}\tilde{H}}}{\partial\alpha}=\frac{-b_{\hat{T}}\beta(1-b_{\hat{T}})}{[(1-\alpha)b_{\hat{T}}+\beta(1-b_{\hat{T}})]^2}<0, \hat{T}\in\{\hat{L},\hat{H}\}$.
(ii) $\frac{\partial b_{\hat{H}\tilde{L}}}{\partial\beta}=\frac{\alpha(1-b_{\hat{H}})}{[\alpha b_{\hat{H}}+(1-\beta)(1-b_{\hat{H}})]^2}>0, \frac{\partial b_{\hat{H}\tilde{H}}}{\partial\beta}=\frac{-(1-\alpha)b_{\hat{H}}(1-b_{\hat{H}}}{[(1-\alpha)b_{\hat{H}}+\beta(1-b_{\hat{H}})]^2}<0, \hat{T}\in\{\hat{L},\hat{H}\}$.
(iii) $\frac{\partial\lambda_{\hat{T}\tilde{L}}}{\partial\alpha}=b_{\hat{T}}\lambda_{\hat{T}}>0, \frac{\partial\lambda_{\hat{T}\tilde{H}}}{\partial\alpha}=-b_{\hat{T}}\lambda_{\hat{T}}<0, \hat{T}\in\{\hat{L},\hat{H}\}$.
(iv) $\frac{\partial\lambda_{\hat{T}\tilde{L}}}{\partial\beta}=-(1-b_{\hat{T}})\lambda_{\hat{T}}<0, \frac{\partial\lambda_{\hat{T}\tilde{H}}}{\partial\beta}=(1-b_{\hat{T}})\lambda_{\hat{T}}>0, \hat{T}\in\{\hat{L},\hat{H}\}$. □

Proof of Proposition 3. The uniqueness of equilibrium patient flow follows directly from Lemma EC.3 with $n=4$. We then show that $\forall\ \alpha,\beta$ s.t. $\alpha\geq 0,\beta\geq 0,\alpha+\beta\leq 1$, we have $f^e_{\hat{L}\tilde{L}}=0$ and $f^e_{\hat{H}\tilde{H}}=1$ in equilibrium. We prove by contradiction. Let $C_{t,l}(f_{\hat{L}\tilde{L}},f_{\hat{L}\tilde{H}},f_{\hat{H}\tilde{L}},f_{\hat{H}\tilde{H}})$ denote the patient cost of a type t patient going to l under patient flow $(f_{\hat{L}\tilde{L}},f_{\hat{L}\tilde{H}},f_{\hat{H}\tilde{L}},f_{\hat{H}\tilde{H}})$, where $t\in\{\hat{L}\tilde{L},\hat{L}\tilde{H},\hat{H}\tilde{L},\hat{H}\tilde{H}\}$ and $l\in\{G,E\}$.

Suppose $\exists\ \alpha,\beta$ s.t. $\alpha\geq 0,\beta\geq 0,\alpha+\beta\leq 1$, and $f^e_{\hat{L}\tilde{L}}>0$, i.e., $\hat{L}\tilde{L}$ patients go to ED directly with a positive probability in equilibrium. This means that

$$
\begin{aligned}
&C_{\hat{L}\tilde{L},E}(f_{\hat{L}\tilde{L}}=0,f_{\hat{L}\tilde{H}}=1,f_{\hat{H}\tilde{L}}=1,f_{\hat{H}\tilde{H}}=1)<C_{\hat{L}\tilde{L},G}(f_{\hat{L}\tilde{L}}=0,f_{\hat{L}\tilde{H}}=1,f_{\hat{H}\tilde{L}}=1,f_{\hat{H}\tilde{H}}=1)\\
&\quad\Leftrightarrow wQ_E(b_{\hat{L}\tilde{L}}\lambda_{\hat{L}\tilde{L}}+\lambda_{\hat{L}\tilde{H}}+\lambda_{\hat{H}})+\hat{p}^*_E<wQ_G+\hat{p}^*_G+b_{\hat{L}\tilde{L}}[wQ_E(b_{\hat{L}\tilde{L}}\lambda_{\hat{L}\tilde{L}}+\lambda_{\hat{L}\tilde{H}}+\lambda_{\hat{H}})+\hat{p}^*_E]\\
&\quad\Rightarrow wQ_E(b_{\hat{L}}\lambda_{\hat{L}}+\lambda_{\hat{H}})+\hat{p}^*_E<wQ_G+\hat{p}^*_G+b_{\hat{L}\tilde{L}}[wQ_E(b_{\hat{L}}\lambda_{\hat{L}}+\lambda_{\hat{H}})+\hat{p}^*_E]\\
&\quad\Rightarrow wQ_E(b_{\hat{L}}\lambda_{\hat{L}}+\lambda_{\hat{H}})+\hat{p}^*_E<wQ_G+\hat{p}^*_G+b_{\hat{L}}[wQ_E(b_{\hat{L}}\lambda_{\hat{L}}+\lambda_{\hat{H}})+\hat{p}^*_E]\\
&\quad\Rightarrow M_E(\lambda_H)<wQ_G+a_G+b_{\hat{L}}M_E(\lambda_H)
\end{aligned}
$$

which contradicts Assumption 1 (i). Similarly, suppose $\exists\ \alpha,\beta$ s.t. $\alpha\geq 0,\beta\geq 0,\alpha+\beta\leq 1$ and $f^e_{\hat{H}\tilde{H}}<1$, i.e., $\hat{H}\tilde{H}$ patients go to GP first with a positive probability in equilibrium. This means that

$$
\begin{aligned}
&C_{\hat{H}\tilde{H},E}(f_{\hat{L}\tilde{L}}=0,f_{\hat{L}\tilde{H}}=0,f_{\hat{H}\tilde{L}}=0,f_{\hat{H}\tilde{H}}=1)>C_{\hat{H}\tilde{H},G}(f_{\hat{L}\tilde{L}}=0,f_{\hat{L}\tilde{H}}=0,f_{\hat{H}\tilde{L}}=0,f_{\hat{H}\tilde{H}}=1)\\
&\quad\Leftrightarrow wQ_E(b_{\hat{L}}\lambda_{\hat{L}}+b_{\hat{H}\tilde{L}}\lambda_{\hat{H}\tilde{L}}+\lambda_{\hat{H}\tilde{H}})+\hat{p}^*_E>wQ_G+\hat{p}^*_G+b_{\hat{H}\tilde{H}}[wQ_E(b_{\hat{L}}\lambda_{\hat{L}}+b_{\hat{H}\tilde{L}}\lambda_{\hat{H}\tilde{L}}+\lambda_{\hat{H}\tilde{H}})+\hat{p}^*_E]\\
&\quad\Rightarrow wQ_E(b_{\hat{L}}\lambda_{\hat{L}}+\lambda_{\hat{H}})+\hat{p}^*_E>wQ_G+\hat{p}^*_G+b_{\hat{H}\tilde{H}}[wQ_E(b_{\hat{L}}\lambda_{\hat{L}}+\lambda_{\hat{H}})+\hat{p}^*_E]\\
&\quad\Rightarrow wQ_E(b_{\hat{L}}\lambda_{\hat{L}}+\lambda_{\hat{H}})+\hat{p}^*_E>wQ_G+\hat{p}^*_G+b_{\hat{H}}[wQ_E(b_{\hat{L}}\lambda_{\hat{L}}+\lambda_{\hat{H}})+\hat{p}^*_E]\\
&\quad\Rightarrow M_E(b_{\hat{L}}\lambda_{\hat{L}}+\lambda_{\hat{H}})>wQ_G+a_G+b_{\hat{H}}M_E(b_{\hat{L}}\lambda_{\hat{L}}+\lambda_{\hat{H}})
\end{aligned}
$$

which contradicts Assumption 1 (ii). Hence, we have $f^e_{\hat{L}\tilde{L}}=0$ and $f^e_{\hat{H}\tilde{H}}=1$ in equilibrium, $\forall\ \alpha,\beta$ s.t. $\alpha\geq 0,\beta\geq 0,\alpha+\beta\leq 1$.

By contrast, the values of $f^e_{\hat{L}\tilde{H}}$ and $f^e_{\hat{H}\tilde{L}}$ depend on α and β. Let $R^e_{a,b}$ denote the equilibrium regime, where $a=f^e_{\hat{L}\tilde{H}},b=f^e_{\hat{H}\tilde{L}}$. In particular, when $f^e_{\hat{L}\tilde{H}}\in\{0,1\}$ and $f^e_{\hat{H}\tilde{L}}\in\{0,1\}$, we have four different pure strategy equilibrium regimes: $R^e_{0,1},R^e_{1,1},R^e_{0,0},R^e_{1,0}$; when $f^e_{\hat{L}\tilde{H}}\in(0,1)$ or $f^e_{\hat{H}\tilde{L}}\in(0,1)$, we have four different mixed strategy equilibrium regimes: $R^e_{(0,1),1},R^e_{(0,1),0},R^e_{0,(0,1)},R^e_{1,(0,1)}$. We characterize the relative position of each regime by characterizing their boundaries.

(1) $R^e_{0,1}$: The pure strategy equilibrium regime $R^e_{0,1}$ is achieved with α and β s.t.

$$
\begin{aligned}
&C_{\hat{L}\tilde{H},E}(f_{\hat{L}\tilde{L}}=0,f_{\hat{L}\tilde{H}}=0,f_{\hat{H}\tilde{L}}=1,f_{\hat{H}\tilde{H}}=1)\geq C_{\hat{L}\tilde{H},G}(f_{\hat{L}\tilde{L}}=0,f_{\hat{L}\tilde{H}}=0,f_{\hat{H}\tilde{L}}=1,f_{\hat{H}\tilde{H}}=1)\\
&\quad\Leftrightarrow(1-b_{\hat{L}\tilde{H}})[wQ_E(b_{\hat{L}}\lambda_{\hat{L}}+\lambda_{\hat{H}})+\hat{p}^*_E]\geq wQ_G+\hat{p}^*_G
\end{aligned}
$$

and

$$C_{\hat{H}\tilde{L},E}(f_{\hat{L}\tilde{L}}=0,f_{\hat{L}\tilde{H}}=0,f_{\hat{H}\tilde{L}}=1,f_{\hat{H}\tilde{H}}=1)\le C_{\hat{H}\tilde{L},G}(f_{\hat{L}\tilde{L}}=0,f_{\hat{L}\tilde{H}}=0,f_{\hat{H}\tilde{L}}=1,f_{\hat{H}\tilde{H}}=1)$$
$$\Leftrightarrow (1-b_{\hat{H}\tilde{L}})[wQ_E(b_{\hat{L}}\lambda_{\hat{L}}+\lambda_{\hat{H}})+\hat{p}_E^*]\le wQ_G+\hat{p}_G^*$$

(2) $R_{1,1}^e$: The pure strategy equilibrium regime $R_{1,1}^e$ is achieved with α and β s.t.

$$C_{\hat{L}\tilde{H},E}(f_{\hat{L}\tilde{L}}=0,f_{\hat{L}\tilde{H}}=1,f_{\hat{H}\tilde{L}}=1,f_{\hat{H}\tilde{H}}=1)\le C_{\hat{L}\tilde{H},G}(f_{\hat{L}\tilde{L}}=0,f_{\hat{L}\tilde{H}}=1,f_{\hat{H}\tilde{L}}=1,f_{\hat{H}\tilde{H}}=1)$$
$$\Leftrightarrow (1-b_{\hat{L}\tilde{H}})[wQ_E((b_{\hat{L}}+\beta(1-b_{\hat{L}}))\lambda_{\hat{L}}+\lambda_{\hat{H}})+\hat{p}_E^*]\le wQ_G+\hat{p}_G^*$$

and

$$C_{\hat{H}\tilde{L},E}(f_{\hat{L}\tilde{L}}=0,f_{\hat{L}\tilde{H}}=1,f_{\hat{H}\tilde{L}}=1,f_{\hat{H}\tilde{H}}=1)\le C_{\hat{H}\tilde{L},G}(f_{\hat{L}\tilde{L}}=0,f_{\hat{L}\tilde{H}}=1,f_{\hat{H}\tilde{L}}=1,f_{\hat{H}\tilde{H}}=1)$$
$$\Leftrightarrow (1-b_{\hat{H}\tilde{L}})[wQ_E((b_{\hat{L}}+\beta(1-b_{\hat{L}}))\lambda_{\hat{L}}+\lambda_{\hat{H}})+\hat{p}_E^*]\le wQ_G+\hat{p}_G^*$$

(3) $R_{0,0}^e$: The pure strategy equilibrium regime $R_{0,0}^e$ is achieved with α and β s.t.

$$C_{\hat{H}\tilde{L},E}(f_{\hat{L}\tilde{L}}=0,f_{\hat{L}\tilde{H}}=0,f_{\hat{H}\tilde{L}}=0,f_{\hat{H}\tilde{H}}=1)\ge C_{\hat{H}\tilde{L},G}(f_{\hat{L}\tilde{L}}=0,f_{\hat{L}\tilde{H}}=0,f_{\hat{H}\tilde{L}}=0,f_{\hat{H}\tilde{H}}=1)$$
$$\Leftrightarrow (1-b_{\hat{H}\tilde{L}})[wQ_E(b_{\hat{L}}\lambda_{\hat{L}}+b_{\hat{H}}\lambda_{\hat{H}}+\beta(1-b_{\hat{H}})\lambda_{\hat{H}})+\hat{p}_E^*]\ge wQ_G+\hat{p}_G^*$$

and

$$C_{\hat{L}\tilde{H},E}(f_{\hat{L}\tilde{L}}=0,f_{\hat{L}\tilde{H}}=0,f_{\hat{H}\tilde{L}}=0,f_{\hat{H}\tilde{H}}=1)\ge C_{\hat{L}\tilde{H},G}(f_{\hat{L}\tilde{L}}=0,f_{\hat{L}\tilde{H}}=0,f_{\hat{H}\tilde{L}}=0,f_{\hat{H}\tilde{H}}=1)$$
$$\Leftrightarrow (1-b_{\hat{L}\tilde{H}})[wQ_E(b_{\hat{L}}\lambda_{\hat{L}}+b_{\hat{H}}\lambda_{\hat{H}}+\beta(1-b_{\hat{H}})\lambda_{\hat{H}})+\hat{p}_E^*]\ge wQ_G+\hat{p}_G^*$$

(4) $R_{1,0}^e$: The pure strategy equilibrium regime $R_{1,0}^e$ is achieved with α and β s.t.

$$C_{\hat{H}\tilde{L},E}(f_{\hat{L}\tilde{L}}=0,f_{\hat{L}\tilde{H}}=1,f_{\hat{H}\tilde{L}}=0,f_{\hat{H}\tilde{H}}=1)\ge C_{\hat{H}\tilde{L},G}(f_{\hat{L}\tilde{L}}=0,f_{\hat{L}\tilde{H}}=1,f_{\hat{H}\tilde{L}}=0,f_{\hat{H}\tilde{H}}=1)$$
$$\Leftrightarrow (1-b_{\hat{H}\tilde{L}})[wQ_E(b_{\hat{L}}\lambda_{\hat{L}}+b_{\hat{H}}\lambda_{\hat{H}}+\beta(1-b_{\hat{L}})\lambda_{\hat{L}}+\beta(1-b_{\hat{H}})\lambda_{\hat{H}})+\hat{p}_E^*]\ge wQ_G+\hat{p}_G^*$$

and

$$C_{\hat{L}\tilde{H},E}(f_{\hat{L}\tilde{L}}=0,f_{\hat{L}\tilde{H}}=1,f_{\hat{H}\tilde{L}}=0,f_{\hat{H}\tilde{H}}=1)\le C_{\hat{L}\tilde{H},G}(f_{\hat{L}\tilde{L}}=0,f_{\hat{L}\tilde{H}}=1,f_{\hat{H}\tilde{L}}=0,f_{\hat{H}\tilde{H}}=1)$$
$$\Leftrightarrow (1-b_{\hat{L}\tilde{H}})[wQ_E(b_{\hat{L}}\lambda_{\hat{L}}+b_{\hat{H}}\lambda_{\hat{H}}+\beta(1-b_{\hat{L}})\lambda_{\hat{L}}+\beta(1-b_{\hat{H}})\lambda_{\hat{H}})+\hat{p}_E^*]\le wQ_G+\hat{p}_G^*$$

(5) $R_{(0,1),1}^e$: The mixed strategy equilibrium regime $R_{(0,1),1}^e$ is achieved with α and β s.t.

$$C_{\hat{L}\tilde{H},E}(f_{\hat{L}\tilde{L}}=0,f_{\hat{L}\tilde{H}}=0,f_{\hat{H}\tilde{L}}=1,f_{\hat{H}\tilde{H}}=1)< C_{\hat{L}\tilde{H},G}(f_{\hat{L}\tilde{L}}=0,f_{\hat{L}\tilde{H}}=0,f_{\hat{H}\tilde{L}}=1,f_{\hat{H}\tilde{H}}=1)$$
$$\Leftrightarrow (1-b_{\hat{L}\tilde{H}})[wQ_E(b_{\hat{L}}\lambda_{\hat{L}}+\lambda_{\hat{H}})+\hat{p}_E^*]< wQ_G+\hat{p}_G^*$$

and

$$C_{\hat{L}\tilde{H},E}(f_{\hat{L}\tilde{L}}=0,f_{\hat{L}\tilde{H}}=1,f_{\hat{H}\tilde{L}}=1,f_{\hat{H}\tilde{H}}=1)>C_{\hat{L}\tilde{H},G}(f_{\hat{L}\tilde{L}}=0,f_{\hat{L}\tilde{H}}=1,f_{\hat{H}\tilde{L}}=1,f_{\hat{H}\tilde{H}}=1)$$
$$\Leftrightarrow (1-b_{\hat{L}\tilde{H}})[wQ_E((b_{\hat{L}}+\beta(1-b_{\hat{L}}))\lambda_{\hat{L}}+\lambda_{\hat{H}})+\hat{p}_E^*]>wQ_G+\hat{p}_G^*$$

(6) $R^e_{(0,1),0}$: The mixed strategy equilibrium regime $R^e_{(0,1),0}$ is achieved with α and β s.t.

$$C_{\hat{L}\tilde{H},E}(f_{\hat{L}\tilde{L}}=0,f_{\hat{L}\tilde{H}}=1,f_{\hat{H}\tilde{L}}=0,f_{\hat{H}\tilde{H}}=1)>C_{\hat{L}\tilde{H},G}(f_{\hat{L}\tilde{L}}=0,f_{\hat{L}\tilde{H}}=1,f_{\hat{H}\tilde{L}}=0,f_{\hat{H}\tilde{H}}=1)$$
$$\Leftrightarrow (1-b_{\hat{L}\tilde{H}})[wQ_E(b_{\hat{L}}\lambda_{\hat{L}}+b_{\hat{H}}\lambda_{\hat{H}}+\beta(1-b_{\hat{L}})\lambda_{\hat{L}}+\beta(1-b_{\hat{H}})\lambda_{\hat{H}})+\hat{p}_E^*]>wQ_G+\hat{p}_G^*$$

and

$$C_{\hat{L}\tilde{H},E}(f_{\hat{L}\tilde{L}}=0,f_{\hat{L}\tilde{H}}=0,f_{\hat{H}\tilde{L}}=0,f_{\hat{H}\tilde{H}}=1)<C_{\hat{L}\tilde{H},G}(f_{\hat{L}\tilde{L}}=0,f_{\hat{L}\tilde{H}}=0,f_{\hat{H}\tilde{L}}=0,f_{\hat{H}\tilde{H}}=1)$$
$$\Leftrightarrow (1-b_{\hat{L}\tilde{H}})[wQ_E(b_{\hat{L}}\lambda_{\hat{L}}+b_{\hat{H}}\lambda_{\hat{H}}+\beta(1-b_{\hat{H}})\lambda_{\hat{H}})+\hat{p}_E^*]<wQ_G+\hat{p}_G^*$$

(7) $R^e_{0,(0,1)}$: The mixed strategy equilibrium regime $R^e_{0,(0,1)}$ is achieved with α and β s.t.

$$C_{\hat{H}\tilde{L},E}(f_{\hat{L}\tilde{L}}=0,f_{\hat{L}\tilde{H}}=0,f_{\hat{H}\tilde{L}}=1,f_{\hat{H}\tilde{H}}=1)>C_{\hat{H}\tilde{L},G}(f_{\hat{L}\tilde{L}}=0,f_{\hat{L}\tilde{H}}=0,f_{\hat{H}\tilde{L}}=1,f_{\hat{H}\tilde{H}}=1)$$
$$\Leftrightarrow (1-b_{\hat{H}\tilde{L}})[wQ_E(b_{\hat{L}}\lambda_{\hat{L}}+\lambda_{\hat{H}})+\hat{p}_E^*]>wQ_G+\hat{p}_G^*$$

and

$$C_{\hat{H}\tilde{L},E}(f_{\hat{L}\tilde{L}}=0,f_{\hat{L}\tilde{H}}=0,f_{\hat{H}\tilde{L}}=0,f_{\hat{H}\tilde{H}}=1)<C_{\hat{H}\tilde{L},G}(f_{\hat{L}\tilde{L}}=0,f_{\hat{L}\tilde{H}}=0,f_{\hat{H}\tilde{L}}=0,f_{\hat{H}\tilde{H}}=1)$$
$$\Leftrightarrow (1-b_{\hat{H}\tilde{L}})[wQ_E(b_{\hat{L}}\lambda_{\hat{L}}+b_{\hat{H}}\lambda_{\hat{H}}+\beta(1-b_{\hat{H}})\lambda_{\hat{H}})+\hat{p}_E^*]<wQ_G+\hat{p}_G^*$$

(8) $R^e_{1,(0,1)}$: The mixed strategy equilibrium regime $R^e_{1,(0,1)}$ is achieved with α and β s.t.

$$C_{\hat{H}\tilde{L},E}(f_{\hat{L}\tilde{L}}=0,f_{\hat{L}\tilde{H}}=1,f_{\hat{H}\tilde{L}}=0,f_{\hat{H}\tilde{H}}=1)<C_{\hat{H}\tilde{L},G}(f_{\hat{L}\tilde{L}}=0,f_{\hat{L}\tilde{H}}=1,f_{\hat{H}\tilde{L}}=0,f_{\hat{H}\tilde{H}}=1)$$
$$\Leftrightarrow (1-b_{\hat{H}\tilde{L}})[wQ_E(b_{\hat{L}}\lambda_{\hat{L}}+b_{\hat{H}}\lambda_{\hat{H}}+\beta(1-b_{\hat{L}})\lambda_{\hat{L}}+\beta(1-b_{\hat{H}})\lambda_{\hat{H}})+\hat{p}_E^*]<wQ_G+\hat{p}_G^*$$

and

$$C_{\hat{H}\tilde{L},E}(f_{\hat{L}\tilde{L}}=0,f_{\hat{L}\tilde{H}}=1,f_{\hat{H}\tilde{L}}=1,f_{\hat{H}\tilde{H}}=1)>C_{\hat{H}\tilde{L},G}(f_{\hat{L}\tilde{L}}=0,f_{\hat{L}\tilde{H}}=1,f_{\hat{H}\tilde{L}}=1,f_{\hat{H}\tilde{H}}=1)$$
$$\Leftrightarrow (1-b_{\hat{H}\tilde{L}})[wQ_E((b_{\hat{L}}+\beta(1-b_{\hat{L}}))\lambda_{\hat{L}}+\lambda_{\hat{H}})+\hat{p}_E^*]>wQ_G+\hat{p}_G^*$$

The relative position of each regime follows from the inequalities above, as shown in Figure 2.

□

Proof of Lemma 4. (i) In equilibrium regime $R^e_{(0,1),1}$ and $R^e_{(0,1),0}$, for a given α and β s.t. we have $f^e_{\hat{L}\tilde{H}} \in (0,1)$, $f^e_{\hat{L}\tilde{H}}$ is determined by solving the following problem:

$$\min_{0<f_{\hat{L}\tilde{H}}<1} \Phi(f_{\hat{L}\tilde{H}}) = \int_0^{\lambda_G} wQ_G dx + \int_0^{\lambda_E} wQ_E(x)dx + \lambda_G \hat{p}^*_G + \lambda_E \hat{p}^*_E \tag{EC.9}$$

where $f^e_{\hat{L}\tilde{H}}$ is given by the following FOC of EC.9:

$$\frac{\partial \Phi(f^e_{\hat{L}\tilde{H}})}{\partial f^e_{\hat{L}\tilde{H}}} = -[(1-\alpha)b_{\hat{L}} + \beta(1-b_{\hat{L}})]\lambda_{\hat{L}}(\hat{p}^*_G + wQ_G) + \beta(1-b_{\hat{L}})\lambda_{\hat{L}}[\hat{p}^*_E + wQ_E(\lambda_E)] = 0 \tag{EC.10}$$

In $R^e_{(0,1),1}$, we have

$$\lambda_G = \lambda_{\hat{L}\tilde{L}} + (1-f^e_{\hat{L}\tilde{H}})\lambda_{\hat{L}\tilde{H}} = \lambda_{\hat{L}} - f^e_{\hat{L}\tilde{H}}[(1-\alpha)b_{\hat{L}} + \beta(1-b_{\hat{L}})]\lambda_{\hat{L}}$$
$$\lambda_E = b_{\hat{L}\tilde{L}}\lambda_{\hat{L}\tilde{L}} + (1-f^e_{\hat{L}\tilde{H}})b_{\hat{L}\tilde{H}}\lambda_{\hat{L}\tilde{H}} + f^e_{\hat{L}\tilde{H}}\lambda_{\hat{L}\tilde{H}} + \lambda_{\hat{H}} = b_{\hat{L}}\lambda_{\hat{L}} + f^e_{\hat{L}\tilde{H}}\beta(1-b_{\hat{L}})\lambda_{\hat{L}} + \lambda_{\hat{H}}$$

By implicit function theorem, we have

$$\frac{\partial f^e_{\hat{L}\tilde{H}}}{\partial \alpha} = -\frac{\partial^2 \Phi(f^e_{\hat{L}\tilde{H}})}{\partial f^e_{\hat{L}\tilde{H}} \partial \alpha} \Big/ \frac{\partial^2 \Phi(f^e_{\hat{L}\tilde{H}})}{\partial f^{e2}_{\hat{L}\tilde{H}}} < 0 \tag{EC.11}$$

as

$$\frac{\partial^2 \Phi(f^e_{\hat{L}\tilde{H}})}{\partial f^{e2}_{\hat{L}\tilde{H}}} = [\beta(1-b_{\hat{L}})\lambda_{\hat{L}}]^2 w \frac{\partial Q_E(\lambda_E)}{\partial \lambda_E} > 0 \tag{EC.12}$$

and

$$\frac{\partial^2 \Phi(f^e_{\hat{L}\tilde{H}})}{\partial f^e_{\hat{L}\tilde{H}} \partial \alpha} = b_{\hat{L}}\lambda_{\hat{L}}(\hat{p}^*_G + wQ_G) > 0 \tag{EC.13}$$

Similarly, we have

$$\frac{\partial f^e_{\hat{L}\tilde{H}}}{\partial \beta} = -\frac{\partial^2 \Phi(f^e_{\hat{L}\tilde{H}})}{\partial f^e_{\hat{L}\tilde{H}} \partial \beta} \Big/ \frac{\partial^2 \Phi(f^e_{\hat{L}\tilde{H}})}{\partial f^{e2}_{\hat{L}\tilde{H}}} < 0 \tag{EC.14}$$

as

$$\begin{aligned}
\frac{\partial^2 \Phi(f^e_{\hat{L}\tilde{H}})}{\partial f^e_{\hat{L}\tilde{H}} \partial \beta} &= (1-b_{\hat{L}})\lambda_{\hat{L}}[-\hat{p}^*_G - wQ_G + \hat{p}^*_E + wQ_E(\lambda_E) + \beta f^e_{\hat{L}\tilde{H}}(1-b_{\hat{L}})\lambda_{\hat{L}} w \frac{\partial Q_E(\lambda_E)}{\partial \lambda_E}] \\
&> (1-b_{\hat{L}})\lambda_{\hat{L}}[-\hat{p}^*_G - wQ_G + \hat{p}^*_E + wQ_E(\lambda_E)] \\
&> (1-b_{\hat{L}})\lambda_{\hat{L}}[-\hat{p}^*_G - wQ_G + \hat{p}^*_E + wQ_E(\lambda_H)] \\
&> (1-b_{\hat{L}})\lambda_{\hat{L}}[-a_G - wQ_G + a_E + wQ_E(\lambda_H)] \\
&> 0
\end{aligned} \tag{EC.15}$$

In $R^e_{(0,1),0}$, we have

$$\lambda_G = \lambda_{\hat{L}\tilde{L}} + \lambda_{\hat{H}\tilde{L}}(1-f^e_{\hat{L}\tilde{H}})\lambda_{\hat{L}\tilde{H}} = \lambda_{\hat{L}} - f^e_{\hat{L}\tilde{H}}[(1-\alpha)b_{\hat{L}} + \beta(1-b_{\hat{L}})]\lambda_{\hat{L}} + [\alpha b_{\hat{H}} + (1-\beta)(1-b_{\hat{H}})]\lambda_{\hat{H}}$$
$$\begin{aligned}
\lambda_E &= b_{\hat{L}\tilde{L}}\lambda_{\hat{L}\tilde{L}} + b_{\hat{H}\tilde{L}}\lambda_{\hat{H}\tilde{L}} + (1-f^e_{\hat{L}\tilde{H}})b_{\hat{L}\tilde{H}}\lambda_{\hat{L}\tilde{H}} + f^e_{\hat{L}\tilde{H}}\lambda_{\hat{L}\tilde{H}} + \lambda_{\hat{H}\tilde{H}} \\
&= b_{\hat{L}}\lambda_{\hat{L}} + b_{\hat{H}}\lambda_{\hat{H}} + \beta[f^e_{\hat{L}\tilde{H}}(1-b_{\hat{L}})\lambda_{\hat{L}} + (1-b_{\hat{H}})\lambda_{\hat{H}}]
\end{aligned}$$

By implicit function theorem, we have

$$\frac{\partial f^e_{\hat{L}\tilde{H}}}{\partial \alpha} = -\frac{\partial^2 \Phi(f^e_{\hat{L}\tilde{H}})}{\partial f^e_{\hat{L}\tilde{H}} \partial \alpha} \Big/ \frac{\partial^2 \Phi(f^e_{\hat{L}\tilde{H}})}{\partial f^{e2}_{\hat{L}\tilde{H}}} < 0 \tag{EC.16}$$

as

$$\frac{\partial^2 \Phi(f^e_{\hat{L}\tilde{H}})}{\partial f^{e2}_{\hat{L}\tilde{H}}} = [\beta(1-b_{\hat{L}})\lambda_{\hat{L}}]^2 w \frac{\partial Q_E(\lambda_E)}{\partial \lambda_E} > 0 \tag{EC.17}$$

and

$$\frac{\partial^2 \Phi(f^e_{\hat{L}\tilde{H}})}{\partial f^e_{\hat{L}\tilde{H}} \partial \alpha} = b_{\hat{L}} \lambda_{\hat{L}} (\hat{p}^*_G + wQ_G) > 0 \tag{EC.18}$$

Similarly, we have

$$\frac{\partial f^e_{\hat{L}\tilde{H}}}{\partial \beta} = -\frac{\partial^2 \Phi(f^e_{\hat{L}\tilde{H}})}{\partial f^e_{\hat{L}\tilde{H}} \partial \beta} \Big/ \frac{\partial^2 \Phi(f^e_{\hat{L}\tilde{H}})}{\partial f^{e2}_{\hat{L}\tilde{H}}} < 0 \tag{EC.19}$$

as

$$\begin{aligned}
\frac{\partial^2 \Phi(f^e_{\hat{L}\tilde{H}})}{\partial f^e_{\hat{L}\tilde{H}} \partial \beta} &= (1-b_{\hat{L}})\lambda_{\hat{L}}[-\hat{p}^*_G - wQ_G + \hat{p}^*_E + wQ_E(\lambda_E) + \beta(f^e_{\hat{L}\tilde{H}}(1-b_{\hat{L}})\lambda_{\hat{L}} + (1-b_{\hat{H}})\lambda_{\hat{H}}) w \frac{\partial Q_E(\lambda_E)}{\partial \lambda_E}] \\
&> (1-b_{\hat{L}})\lambda_{\hat{L}}[-\hat{p}^*_G - wQ_G + \hat{p}^*_E + wQ_E(\lambda_E)] \\
&> (1-b_{\hat{L}})\lambda_{\hat{L}}[-a_G - wQ_G + a_E + wQ_E(\lambda_H)] \\
&> 0
\end{aligned} \tag{EC.20}$$

(ii) In equilibrium regime $R^e_{0,(0,1)}$ and $R^e_{1,(0,1)}$, for a given α and β s.t. we have $f^e_{\hat{H}\tilde{L}} \in (0,1)$, $f^e_{\hat{H}\tilde{L}}$ is determined by solving the following problem:

$$\min_{0<f_{\hat{H}\tilde{L}}<1} \Phi(f_{\hat{H}\tilde{L}}) = \int_0^{\lambda_G} wQ_G dx + \int_0^{\lambda_E} wQ_E(x)dx + \lambda_G \hat{p}^*_G + \lambda_E \hat{p}^*_E \tag{EC.21}$$

where $f^e_{\hat{H}\tilde{L}}$ is given by the following FOC of EC.21:

$$\frac{\partial \Phi(f^e_{\hat{L}\tilde{H}})}{\partial f^e_{\hat{L}\tilde{H}}} = -[\alpha b_{\hat{H}} + (1-\beta)(1-b_{\hat{H}})]\lambda_{\hat{H}}(\hat{p}^*_G + wQ_G) + (1-\beta)(1-b_{\hat{H}})\lambda_{\hat{H}}[\hat{p}^*_E + wQ_E(\lambda_E)] = 0 \tag{EC.22}$$

In $R^e_{0,(0,1)}$, we have

$$\begin{aligned}
\lambda_G &= \lambda_{\hat{L}} + (1 - f^e_{\hat{H}\tilde{L}})\lambda_{\hat{H}\tilde{L}} = \lambda_{\hat{L}} + (1 - f^e_{\hat{H}\tilde{L}})[\alpha b_{\hat{H}} + (1-\beta)(1-b_{\hat{H}})]\lambda_{\hat{H}} \\
\lambda_E &= b_{\hat{L}}\lambda_{\hat{L}} + (1 - f^e_{\hat{H}\tilde{L}}) b_{\hat{H}\tilde{L}} \lambda_{\hat{H}\tilde{L}} + f^e_{\hat{H}\tilde{L}} \lambda_{\hat{H}\tilde{L}} + \lambda_{\hat{H}\tilde{H}} \\
&= b_{\hat{L}}\lambda_{\hat{L}} + f^e_{\hat{H}\tilde{L}}(1-\beta)(1-b_{\hat{H}})\lambda_{\hat{H}} + [b_{\hat{H}} + \beta(1-b_{\hat{H}})]\lambda_{\hat{H}}
\end{aligned}$$

By implicit function theorem, we have

$$\frac{\partial f^e_{\hat{H}\tilde{L}}}{\partial \alpha} = -\frac{\partial^2 \Phi(f^e_{\hat{H}\tilde{L}})}{\partial f^e_{\hat{H}\tilde{L}} \partial \alpha} \Big/ \frac{\partial^2 \Phi(f^e_{\hat{H}\tilde{L}})}{\partial f^{e2}_{\hat{H}\tilde{L}}} > 0 \tag{EC.23}$$

as

$$\frac{\partial^2\Phi(f^e_{\hat{H}\tilde{L}})}{\partial f^{e2}_{\hat{H}\tilde{L}}} = [(1-\beta)(1-b_{\hat{H}})\lambda_{\hat{H}}]^2 w\frac{\partial Q_E(\lambda_E)}{\partial \lambda_E} > 0 \tag{EC.24}$$

and

$$\frac{\partial^2\Phi(f^e_{\hat{H}\tilde{L}})}{\partial f^e_{\hat{H}\tilde{L}}\partial\alpha} = -b_{\hat{H}}\lambda_{\hat{H}}(\hat{p}^*_G + wQ_G) < 0 \tag{EC.25}$$

In $R^e_{1,(0,1)}$, we have

$$\begin{aligned}
\lambda_G &= \lambda_{\hat{L}\tilde{L}} + (1-f^e_{\hat{H}\tilde{L}})\lambda_{\hat{H}\tilde{L}} = [\alpha b_{\hat{L}} + (1-\beta)(1-b_{\hat{L}}]\lambda_{\hat{L}} + (1-f^e_{\hat{H}\tilde{L}})[\alpha b_{\hat{H}} + (1-\beta)(1-b_{\hat{H}})]\lambda_{\hat{H}} \\
\lambda_E &= b_{\hat{L}\tilde{L}}\lambda_{\hat{L}\tilde{L}} + (1-f^e_{\hat{H}\tilde{L}})b_{\hat{H}\tilde{L}}\lambda_{\hat{H}\tilde{L}} + f^e_{\hat{H}\tilde{L}}\lambda_{\hat{H}\tilde{L}} + \lambda_{\hat{L}\tilde{H}} + \lambda_{\hat{H}\tilde{H}} \\
&= [b_{\hat{L}} + \beta(1-b_{\hat{L}})]\lambda_{\hat{L}} + f^e_{\hat{H}\tilde{L}}(1-\beta)(1-b_{\hat{H}})\lambda_{\hat{H}} + [b_{\hat{H}} + \beta(1-b_{\hat{H}})]\lambda_{\hat{H}}
\end{aligned}$$

By implicit function theorem, we have

$$\frac{\partial f^e_{\hat{H}\tilde{L}}}{\partial\alpha} = -\frac{\partial^2\Phi(f^e_{\hat{H}\tilde{L}})}{\partial f^e_{\hat{H}\tilde{L}}\partial\alpha} \Big/ \frac{\partial^2\Phi(f^e_{\hat{H}\tilde{L}})}{\partial f^{e2}_{\hat{H}\tilde{L}}} > 0 \tag{EC.26}$$

as

$$\frac{\partial^2\Phi(f^e_{\hat{H}\tilde{L}})}{\partial f^{e2}_{\hat{H}\tilde{L}}} = [(1-\beta)(1-b_{\hat{H}})\lambda_{\hat{H}}]^2 w\frac{\partial Q_E(\lambda_E)}{\partial \lambda_E} > 0 \tag{EC.27}$$

and

$$\frac{\partial^2\Phi(f^e_{\hat{H}\tilde{L}})}{\partial f^e_{\hat{H}\tilde{L}}\partial\alpha} = -b_{\hat{H}}\lambda_{\hat{H}}(\hat{p}^*_G + wQ_G) < 0 \tag{EC.28}$$

□

EC.2.4. Proofs for Section 6

Proof of Proposition 4. We have the equilibrium social cost:

$$C^e_s(\alpha,\beta) = \lambda_G wQ_G + \lambda_E wQ_E(\lambda_E) + \lambda_G a_G + \lambda_E a_E \tag{EC.29}$$

(i) In $R^e_{0,1}$, we have $\lambda_G = \lambda_{\hat{L}}, \lambda_E = b_{\hat{L}}\lambda_{\hat{L}} + \lambda_{\hat{H}}$, and therefore $\frac{\partial C^e_s(\alpha,\beta)}{\partial\alpha} = \frac{\partial C^e_s(\alpha,\beta)}{\partial\beta} = 0$.

(ii) In equilibrium regime $R^e_{1,1}$, we have $\lambda_G = \lambda_{\hat{L}\tilde{L}} = [\alpha b_{\hat{L}} + (1-\beta)(1-b_{\hat{L}})]\lambda_{\hat{L}}, \lambda_E = b_{\hat{L}\tilde{L}}\lambda_{\hat{L}\tilde{L}} + \lambda_{\hat{L}\tilde{H}} + \lambda_{\hat{H}} = [b_{\hat{L}} + \beta(1-b_{\hat{L}})]\lambda_{\hat{L}} + \lambda_{\hat{H}}$. We then have $\frac{\partial\lambda_G}{\partial\alpha} = b_{\hat{L}}\lambda_{\hat{L}}, \frac{\partial\lambda_G}{\partial\beta} = -(1-b_{\hat{L}})\lambda_{\hat{L}}, \frac{\partial\lambda_E}{\partial\alpha} = 0, \frac{\partial\lambda_E}{\partial\beta} = (1-b_{\hat{L}})\lambda_{\hat{L}}$.

In equilibrium regime $R^e_{0,0}$, we have $\lambda_G = \lambda_{\hat{L}} + \lambda_{\hat{H}\tilde{L}} = \lambda_{\hat{L}} + [\alpha b_{\hat{H}} + (1-\beta)(1-b_{\hat{H}})]\lambda_{\hat{H}}, \lambda_E = b_{\hat{L}}\lambda_{\hat{L}} + b_{\hat{H}\tilde{L}}\lambda_{\hat{H}\tilde{L}} + \lambda_{\hat{H}\tilde{H}} = b_{\hat{L}}\lambda_{\hat{L}} + b_{\hat{H}}\lambda_{\hat{H}} + \beta(1-b_{\hat{H}})\lambda_{\hat{H}}$. We then have $\frac{\partial\lambda_G}{\partial\alpha} = b_{\hat{H}}\lambda_{\hat{H}}, \frac{\partial\lambda_G}{\partial\beta} = -(1-b_{\hat{H}})\lambda_{\hat{H}}, \frac{\partial\lambda_E}{\partial\alpha} = 0, \frac{\partial\lambda_E}{\partial\beta} = (1-b_{\hat{H}})\lambda_{\hat{H}}$.

In equilibrium regime $R^e_{1,0}$, we have $\lambda_G = \lambda_{\hat{L}\tilde{L}} + \lambda_{\hat{H}\tilde{L}} = \alpha(b_{\hat{L}}\lambda_{\hat{L}} + b_{\hat{H}}\lambda_{\hat{H}}) + (1-\beta)[(1-b_{\hat{L}})\lambda_{\hat{L}} + (1-b_{\hat{H}})\lambda_{\hat{H}}], \lambda_E = b_{\hat{L}\tilde{L}}\lambda_{\hat{L}\tilde{L}} + b_{\hat{H}\tilde{L}}\lambda_{\hat{H}\tilde{L}} + \lambda_{\hat{L}\tilde{H}} + \lambda_{\hat{H}\tilde{H}} = b_{\hat{L}}\lambda_{\hat{L}} + b_{\hat{H}}\lambda_{\hat{H}} + \beta[(1-b_{\hat{L}})\lambda_{\hat{L}} + (1-b_{\hat{H}})\lambda_{\hat{H}}]$. We then have $\frac{\partial\lambda_G}{\partial\alpha} = b_{\hat{L}}\lambda_{\hat{L}} + b_{\hat{H}}\lambda_{\hat{H}}, \frac{\partial\lambda_G}{\partial\beta} = -[(1-b_{\hat{L}})\lambda_{\hat{L}} + (1-b_{\hat{H}})\lambda_{\hat{H}}], \frac{\partial\lambda_E}{\partial\alpha} = 0, \frac{\partial\lambda_E}{\partial\beta} = (1-b_{\hat{L}})\lambda_{\hat{L}} + (1-b_{\hat{H}})\lambda_{\hat{H}}$.

Hence, in $R^e_{1,1}, R^e_{0,0}$ and $R^e_{1,0}$, we have $\frac{\partial \lambda_G}{\partial \alpha} > 0, \frac{\partial \lambda_E}{\partial \alpha} = 0$, and therefore $\frac{\partial C^e_s(\alpha,\beta)}{\partial \alpha} > 0$. On the other hand, we have $\frac{\partial \lambda_G}{\partial \beta} < 0, \frac{\partial \lambda_E}{\partial \beta} > 0$, and $\frac{\partial \lambda_G}{\partial \beta} + \frac{\partial \lambda_E}{\partial \beta} = 0$. Since an arrival to an ED is more costly than an arrival to a GP, i.e., $a_G + wQ_G < M_E(\lambda_H)$ as implied by Assumption 1 (i), we have $\frac{\partial C^e_s(\alpha,\beta)}{\partial \beta} > 0$.

(iii) In $R^e_{(0,1),1}$ and $R^e_{(0,1),0}$, we have

$$\begin{aligned}\frac{\partial C^e_s(\alpha,\beta)}{\partial \beta} &= \frac{\partial C^e_s(\alpha,\beta,f^e_{\hat{L}\tilde{H}})}{\partial \beta} + \frac{\partial C^e_s(\alpha,\beta,f^e_{\hat{L}\tilde{H}})}{\partial f^e_{\hat{L}\tilde{H}}}\frac{\partial f^e_{\hat{L}\tilde{H}}}{\partial \beta} \\ &= [\frac{\partial C^e_s(\alpha,\beta,f^e_{\hat{L}\tilde{H}})}{\partial \beta}\frac{\partial^2 \Phi(f^e_{\hat{L}\tilde{H}})}{\partial f^{e2}_{\hat{L}\tilde{H}}} - \frac{\partial C^e_s(\alpha,\beta,f^e_{\hat{L}\tilde{H}})}{\partial f^e_{\hat{L}\tilde{H}}}\frac{\partial^2 \Phi(f^e_{\hat{L}\tilde{H}})}{\partial f^e_{\hat{L}\tilde{H}}\partial \beta}]/\frac{\partial^2 \Phi(f^e_{\hat{L}\tilde{H}})}{\partial f^{e2}_{\hat{L}\tilde{H}}}\end{aligned} \tag{EC.30}$$

In $R^e_{(0,1),1}$, we have

$$\frac{\partial C^e_s(\alpha,\beta,f^e_{\hat{L}\tilde{H}})}{\partial \beta} = f^e_{\hat{L}\tilde{H}}(1-b_{\hat{L}})\lambda_{\hat{L}}[a_E + wQ_E(\lambda_E) + \lambda_E w\frac{\partial Q_E(\lambda_E)}{\partial \lambda_E} - a_G - wQ_G] > 0 \tag{EC.31}$$

and

$$\begin{aligned}&\frac{\partial C^e_s(\alpha,\beta,f^e_{\hat{L}\tilde{H}})}{\partial f^e_{\hat{L}\tilde{H}}} \\ &= -[(1-\alpha)b_{\hat{L}} + \beta(1-b_{\hat{L}})]\lambda_{\hat{L}}(a_G + wQ_G) + \beta(1-b_{\hat{L}})\lambda_{\hat{L}}[a_E + wQ_E(\lambda_E) + \lambda_E w\frac{\partial Q_E(\lambda_E)}{\partial \lambda_E}] \\ &= \beta(1-b_{\hat{L}})\lambda_{\hat{L}}[\lambda_E w\frac{\partial Q_E(\lambda_E)}{\partial \lambda_E} - [\lambda_H w\frac{\partial Q_E(\lambda_E)}{\partial \lambda_E}\Big|_{\lambda_H} + wQ_E(\lambda_H) - wQ_E(b_{\hat{L}}\lambda_{\hat{L}} + \lambda_{\hat{H}})]^+] \\ &> \beta(1-b_{\hat{L}})\lambda_{\hat{L}}[\lambda_E w\frac{\partial Q_E(\lambda_E)}{\partial \lambda_E} - \lambda_H w\frac{\partial Q_E(\lambda_E)}{\partial \lambda_E}\Big|_{\lambda_H}] \\ &> 0\end{aligned} \tag{EC.32}$$

where the second equality is given by Proposition 2 and EC.10. We then have

$$\frac{\partial C^e_s(\alpha,\beta,f^e_{\hat{L}\tilde{H}})}{\partial \beta}\frac{\partial^2 \Phi(f^e_{\hat{L}\tilde{H}})}{\partial f^{e2}_{\hat{L}\tilde{H}}} - \frac{\partial C^e_s(\alpha,\beta,f^e_{\hat{L}\tilde{H}})}{\partial f^e_{\hat{L}\tilde{H}}}\frac{\partial^2 \Phi(f^e_{\hat{L}\tilde{H}})}{\partial f^e_{\hat{L}\tilde{H}}\partial \beta} = \beta[(1-b_{\hat{L}})\lambda_{\hat{L}}]^2(C_1C_2 - C_3C_4) \tag{EC.33}$$

where

$$\begin{aligned}C_1 &= f^e_{\hat{L}\tilde{H}}\beta(1-b_{\hat{L}})\lambda_{\hat{L}} w\frac{\partial Q_E(\lambda_E)}{\partial \lambda_E} > 0 \\ C_2 &= a_E + wQ_E(\lambda_E) + \lambda_E w\frac{\partial Q_E(\lambda_E)}{\partial \lambda_E} - a_G - wQ_G > 0 \\ C_3 &= \lambda_E w\frac{\partial Q_E(\lambda_E)}{\partial \lambda_E} - [\lambda_H w\frac{\partial Q_E(\lambda_E)}{\partial \lambda_E}\Big|_{\lambda_H} + wQ_E(\lambda_H) - wQ_E(b_{\hat{L}}\lambda_{\hat{L}} + \lambda_{\hat{H}})]^+ > 0 \\ C_4 &= a_E + [\lambda_H w\frac{\partial Q_E(\lambda_E)}{\partial \lambda_E}\Big|_{\lambda_H} + wQ_E(\lambda_H) - wQ_E(b_{\hat{L}}\lambda_{\hat{L}} + \lambda_{\hat{H}})]^+ + wQ_E(\lambda_E) \\ &\quad + f^e_{\hat{L}\tilde{H}}\beta(1-b_{\hat{L}})\lambda_{\hat{L}} w\frac{\partial Q_E(\lambda_E)}{\partial \lambda_E} - a_G - wQ_G > 0\end{aligned}$$

Define

$$C_5 = (b_{\hat{L}}\lambda_{\hat{L}} + \lambda_{\hat{H}})w\frac{\partial Q_E(\lambda_E)}{\partial \lambda_E} - [\lambda_H w\frac{\partial Q_E(\lambda_E)}{\partial \lambda_E}\Big|_{\lambda_H} + wQ_E(\lambda_H) - wQ_E(b_{\hat{L}}\lambda_{\hat{L}} + \lambda_{\hat{H}})]^+ > 0$$

We then have $C_3 = C_1 + C_5$ and $C_4 = C_2 - C_5$. Hence we have $C_1C_2 - C_3C_4 = C_1C_2 - (C_1 + C_5)(C_2 - C_5) = (C_1 - C_2 + C_5)C_5$, with

$$C_1 - C_2 + C_5 = a_G + wQ_G - a_E - wQ_E(\lambda_E) - [\lambda_H w \frac{\partial Q_E(\lambda_E)}{\partial \lambda_E}\Big|_{\lambda_H} + wQ_E(\lambda_H) - wQ_E(b_{\hat{L}}\lambda_{\hat{L}} + \lambda_{\hat{H}})]^+ < 0$$

Hence, we have $C_1C_2 - C_3C_4 < 0$. Since we have $\frac{\partial^2 \Phi(f^e_{\hat{L}\tilde{H}})}{\partial f^{e2}_{\hat{H}\tilde{L}}} > 0$ by EC.12, we have $\frac{\partial C^e_s(\alpha,\beta)}{\partial \beta} < 0$ in $R^e_{(0,1),1}$.

In $R^e_{(0,1),0}$, we have

$$\frac{\partial C^e_s(\alpha,\beta,f^e_{\hat{L}\tilde{H}})}{\partial \beta} = [f^e_{\hat{L}\tilde{H}}(1-b_{\hat{L}})\lambda_{\hat{L}} + (1-b_{\hat{H}})\lambda_{\hat{H}}][a_E + wQ_E(\lambda_E) + \lambda_E w \frac{\partial Q_E(\lambda_E)}{\partial \lambda_E} - a_G - wQ_G] > 0 \tag{EC.34}$$

and

$$\begin{aligned}
&\frac{\partial C^e_s(\alpha,\beta,f^e_{\hat{L}\tilde{H}})}{\partial f^e_{\hat{L}\tilde{H}}} \\
&= -[(1-\alpha)b_{\hat{L}} + \beta(1-b_{\hat{L}})]\lambda_{\hat{L}}(a_G + wQ_G) + \beta(1-b_{\hat{L}})\lambda_{\hat{L}}[a_E + wQ_E(\lambda_E) + \lambda_E w \frac{\partial Q_E(\lambda_E)}{\partial \lambda_E}] \\
&= \beta(1-b_{\hat{L}})\lambda_{\hat{L}}[\lambda_E w \frac{\partial Q_E(\lambda_E)}{\partial \lambda_E} - [\lambda_H w \frac{\partial Q_E(\lambda_E)}{\partial \lambda_E}\Big|_{\lambda_H} + wQ_E(\lambda_H) - wQ_E(b_{\hat{L}}\lambda_{\hat{L}} + \lambda_{\hat{H}})]^+] \\
&> 0
\end{aligned} \tag{EC.35}$$

where the second equality is given by Proposition 2 and EC.10. We then have

$$\frac{\partial C^e_s(\alpha,\beta,f^e_{\hat{L}\tilde{H}})}{\partial \beta}\frac{\partial^2 \Phi(f^e_{\hat{L}\tilde{H}})}{\partial f^{e2}_{\hat{L}\tilde{H}}} - \frac{\partial C^e_s(\alpha,\beta,f^e_{\hat{L}\tilde{H}})}{\partial f^e_{\hat{L}\tilde{H}}}\frac{\partial^2 \Phi(f^e_{\hat{L}\tilde{H}})}{\partial f^e_{\hat{L}\tilde{H}}\partial \beta} = \beta[(1-b_{\hat{L}})\lambda_{\hat{L}}]^2(C_1C_2 - C_3C_4) \tag{EC.36}$$

where

$$\begin{aligned}
C_1 &= \beta[f^e_{\hat{L}\tilde{H}}(1-b_{\hat{L}})\lambda_{\hat{L}} + (1-b_{\hat{H}})\lambda_{\hat{H}}]w\frac{\partial Q_E(\lambda_E)}{\partial \lambda_E} > 0 \\
C_2 &= a_E + wQ_E(\lambda_E) + \lambda_E w \frac{\partial Q_E(\lambda_E)}{\partial \lambda_E} - a_G - wQ_G > 0 \\
C_3 &= \lambda_E w \frac{\partial Q_E(\lambda_E)}{\partial \lambda_E} - [\lambda_H w \frac{\partial Q_E(\lambda_E)}{\partial \lambda_E}\Big|_{\lambda_H} + wQ_E(\lambda_H) - wQ_E(b_{\hat{L}}\lambda_{\hat{L}} + \lambda_{\hat{H}})]^+ > 0 \\
C_4 &= a_E + [\lambda_H w \frac{\partial Q_E(\lambda_E)}{\partial \lambda_E}\Big|_{\lambda_H} + wQ_E(\lambda_H) - wQ_E(b_{\hat{L}}\lambda_{\hat{L}} + \lambda_{\hat{H}})]^+ + wQ_E(\lambda_E) \\
&\quad + \beta[f^e_{\hat{L}\tilde{H}}(1-b_{\hat{L}})\lambda_{\hat{L}} + (1-b_{\hat{H}})\lambda_{\hat{H}}]w\frac{\partial Q_E(\lambda_E)}{\partial \lambda_E} - a_G - wQ_G > 0
\end{aligned}$$

Define

$$C_5 = \lambda_H w \frac{\partial Q_E(\lambda_E)}{\partial \lambda_E} - [\lambda_H w \frac{\partial Q_E(\lambda_E)}{\partial \lambda_E}\Big|_{\lambda_H} + wQ_E(\lambda_H) - wQ_E(b_{\hat{L}}\lambda_{\hat{L}} + \lambda_{\hat{H}})]^+ > 0$$

We then have $C_3 = C_1 + C_5$ and $C_4 = C_2 - C_5$. Hence we have $C_1C_2 - C_3C_4 = C_1C_2 - (C_1 + C_5)(C_2 - C_5) = (C_1 - C_2 + C_5)C_5$, with

$$C_1 - C_2 + C_5 = a_G + wQ_G - a_E - wQ_E(\lambda_E) - [\lambda_H w \frac{\partial Q_E(\lambda_E)}{\partial \lambda_E}\Big|_{\lambda_H} + wQ_E(\lambda_H) - wQ_E(b_{\hat{L}}\lambda_{\hat{L}} + \lambda_{\hat{H}})]^+ < 0$$

Hence, we have $C_1C_2 - C_3C_4 < 0$. Since we have $\frac{\partial^2 \Phi(f^e_{\hat{L}\tilde{H}})}{\partial f^{e2}_{\hat{H}\tilde{L}}} > 0$ by EC.17, we have $\frac{\partial C^e_s(\alpha,\beta)}{\partial \beta} < 0$ in $R^e_{(0,1),0}$.

(iv) In $R^e_{0,(0,1)}$ and $R^e_{1,(0,1)}$, we have

$$\begin{aligned}\frac{\partial C^e_s(\alpha,\beta)}{\partial \alpha} &= \frac{\partial C^e_s(\alpha,\beta,f^e_{\hat{H}\tilde{L}})}{\partial \alpha} + \frac{\partial C^e_s(\alpha,\beta,f^e_{\hat{H}\tilde{L}})}{\partial f^e_{\hat{H}\tilde{L}}}\frac{\partial f^e_{\hat{H}\tilde{L}}}{\partial \alpha} \\ &= [\frac{\partial C^e_s(\alpha,\beta,f^e_{\hat{H}\tilde{L}})}{\partial \alpha}\frac{\partial^2 \Phi(f^e_{\hat{H}\tilde{L}})}{\partial f^{e2}_{\hat{H}\tilde{L}}} - \frac{\partial C^e_s(\alpha,\beta,f^e_{\hat{H}\tilde{L}})}{\partial f^e_{\hat{H}\tilde{L}}}\frac{\partial^2 \Phi(f^e_{\hat{H}\tilde{L}})}{\partial f^e_{\hat{H}\tilde{L}}\partial \alpha}]/\frac{\partial^2 \Phi(f^e_{\hat{H}\tilde{L}})}{\partial f^{e2}_{\hat{H}\tilde{L}}}\end{aligned} \tag{EC.37}$$

In $R^e_{0,(0,1)}$, we have

$$\frac{\partial C^e_s(\alpha,\beta,f^e_{\hat{H}\tilde{L}})}{\partial \alpha} = (1 - f^e_{\hat{H}\tilde{L}})b_{\hat{H}}\lambda_{\hat{H}}[a_G + wQ_G] > 0 \tag{EC.38}$$

and

$$\begin{aligned}\frac{\partial C^e_s(\alpha,\beta,f^e_{\hat{H}\tilde{L}})}{\partial f^e_{\hat{H}\tilde{L}}} &= -[\alpha b_{\hat{H}} + (1-\beta)(1-b_{\hat{H}})]\lambda_{\hat{H}}(a_G + wQ_G) \\ &\quad + (1-\beta)(1-b_{\hat{H}})\lambda_{\hat{H}}[a_E + wQ_E(\lambda_E) + \lambda_E w\frac{\partial Q_E(\lambda_E)}{\partial \lambda_E}] \\ &= (1-\beta)(1-b_{\hat{H}})\lambda_{\hat{H}}[\lambda_E w\frac{\partial Q_E(\lambda_E)}{\partial \lambda_E} \\ &\quad - [\lambda_H w\frac{\partial Q_E(\lambda_E)}{\partial \lambda_E}\Big|_{\lambda_H} + wQ_E(\lambda_H) - wQ_E(b_{\hat{L}}\lambda_{\hat{L}} + \lambda_{\hat{H}})]^+] \\ &> 0\end{aligned} \tag{EC.39}$$

where the second equality is given by Proposition 2 and EC.22. Since we have $\frac{\partial^2 \Phi(f^e_{\hat{H}\tilde{L}})}{\partial f^{e2}_{\hat{H}\tilde{L}}} > 0$ by EC.24 and $\frac{\partial^2 \Phi(f^e_{\hat{H}\tilde{L}})}{\partial f^e_{\hat{H}\tilde{L}}\partial \alpha} < 0$ by EC.25, we have $\frac{\partial C^e_s(\alpha,\beta)}{\partial \alpha} > 0$ in $R^e_{0,(0,1)}$.

In $R^e_{1,(0,1)}$, we have

$$\frac{\partial C^e_s(\alpha,\beta,f^e_{\hat{H}\tilde{L}})}{\partial \alpha} = [b_{\hat{L}}\lambda_{\hat{L}} + (1 - f^e_{\hat{H}\tilde{L}})b_{\hat{H}}\lambda_{\hat{H}}][a_G + wQ_G] > 0 \tag{EC.40}$$

and

$$\begin{aligned}\frac{\partial C^e_s(\alpha,\beta,f^e_{\hat{H}\tilde{L}})}{\partial f^e_{\hat{H}\tilde{L}}} &= -[\alpha b_{\hat{H}} + (1-\beta)(1-b_{\hat{H}})]\lambda_{\hat{H}}(a_G + wQ_G) \\ &\quad + (1-\beta)(1-b_{\hat{H}})\lambda_{\hat{H}}[a_E + wQ_E(\lambda_E) + \lambda_E w\frac{\partial Q_E(\lambda_E)}{\partial \lambda_E}] \\ &= (1-\beta)(1-b_{\hat{H}})\lambda_{\hat{H}}[\lambda_E w\frac{\partial Q_E(\lambda_E)}{\partial \lambda_E} \\ &\quad - [\lambda_H w\frac{\partial Q_E(\lambda_E)}{\partial \lambda_E}\Big|_{\lambda_H} + wQ_E(\lambda_H) - wQ_E(b_{\hat{L}}\lambda_{\hat{L}} + \lambda_{\hat{H}})]^+] \\ &> 0\end{aligned} \tag{EC.41}$$

where the second equality is given by Proposition 2 and EC.22. Since we have $\frac{\partial^2\Phi(f^e_{\hat{H}\tilde{L}})}{\partial f^{e2}_{\hat{H}\tilde{L}}} > 0$ by EC.27 and $\frac{\partial^2\Phi(f^e_{\hat{H}\tilde{L}})}{\partial f^e_{\hat{H}\tilde{L}}\partial\alpha} < 0$ by EC.28, we have $\frac{\partial C^e_s(\alpha,\beta)}{\partial\alpha} > 0$ in $R^e_{0,(0,1)}$. □

Proof of Proposition 5. We have $R_{(0,1),1}$ on the left of and being adjunct to $R_{0,1}$ by Proposition 3. In addition, we have $C^e_s(\alpha,\beta) = C_s(f_{\hat{L}} = 0, f_{\hat{H}} = 1)$ in $R_{0,1}$, and $\frac{\partial C^e_s(\alpha,\beta)}{\partial\beta} < 0$ in $R^e_{(0,1),1}$ by Proposition 4 (iii). Hence, it follows that $\exists\ \alpha,\beta$ in $R^e_{(0,1),1}$ s.t. $C^e_s(\alpha,\beta) > C_s(f_{\hat{L}} = 0, f_{\hat{H}} = 1)$. □

EC.2.5. Proofs for Section 7

Proof of Proposition 6. We first prove $f^*_{\hat{L}\tilde{L}} = 0, f^*_{\hat{H}\tilde{H}} = 1, \forall\ \alpha,\beta$ s.t. $\alpha \geq 0, \beta \geq 0, \alpha+\beta \leq 1$.

Suppose $f_{\hat{L}\tilde{L}} \in [0,1], f_{\hat{L}\tilde{H}} = 1, f_{\hat{H}\tilde{L}} = 1, f_{\hat{H}\tilde{H}} = 1$. Then we have

$$\lambda_G = (1-f_{\hat{L}\tilde{L}})\lambda_{\hat{L}\tilde{L}} = (1-f_{\hat{L}\tilde{L}})[\alpha b_{\hat{L}} + (1-\beta)(1-b_{\hat{L}})]\lambda_{\hat{L}}$$
$$\lambda_E = (1-f_{\hat{L}\tilde{L}})b_{\hat{L}\tilde{L}}\lambda_{\hat{L}\tilde{L}} + f_{\hat{L}\tilde{L}}\lambda_{\hat{L}\tilde{L}} + \lambda_{\hat{L}\tilde{H}} + \lambda_{\hat{H}} = [b_{\hat{L}} + \beta(1-b_{\hat{L}})]\lambda_{\hat{L}} + f_{\hat{L}\tilde{L}}(1-\beta)(1-b_{\hat{L}})\lambda_{\hat{L}} + \lambda_{\hat{H}}$$

Hence we have

$$\begin{aligned}
\frac{\partial C_s(f_{\hat{L}\tilde{L}})}{\partial f_{\hat{L}\tilde{L}}} &= -[\alpha b_{\hat{L}} + (1-\beta)(1-b_{\hat{L}})]\lambda_{\hat{L}}(a_G + wQ_G) \\
&\quad + (1-\beta)(1-b_{\hat{L}})\lambda_{\hat{L}}[a_E + wQ_E(\lambda_E) + \lambda_E w\frac{\partial Q_E(\lambda_E)}{\partial\lambda_E}] \\
&= [\alpha b_{\hat{L}} + (1-\beta)(1-b_{\hat{L}})]\lambda_{\hat{L}}[(1-b_{\hat{L}\tilde{L}})(a_E + wQ_E(\lambda_E) + \lambda_E w\frac{\partial Q_E(\lambda_E)}{\partial\lambda_E}) - (a_G + wQ_G)] \\
&\geq [\alpha b_{\hat{L}} + (1-\beta)(1-b_{\hat{L}})]\lambda_{\hat{L}}[(1-b_{\hat{L}\tilde{L}})M_E(\lambda_H) - (a_G + wQ_G)] \\
&\geq [\alpha b_{\hat{L}} + (1-\beta)(1-b_{\hat{L}})]\lambda_{\hat{L}}[(1-b_{\hat{L}})M_E(\lambda_H) - (a_G + wQ_G)] \\
&\geq 0
\end{aligned}$$

Hence we have $f^*_{\hat{L}\tilde{L}} = 0$.

On the other hand, suppose $f_{\hat{L}\tilde{L}} = 0, f_{\hat{L}\tilde{H}} = 0, f_{\hat{H}\tilde{L}} = 0, f_{\hat{H}\tilde{H}} = \in [0,1]$. Then we have

$$\lambda_G = \lambda_{\hat{L}} + \lambda_{\hat{H}\tilde{L}} + (1-f_{\hat{H}\tilde{H}})\lambda_{\hat{H}\tilde{H}} = \lambda_{\hat{L}} + \lambda_{\hat{H}} - f_{\hat{H}\tilde{H}}[(1-\alpha)b_{\hat{H}} + \beta(1-b_{\hat{H}})]\lambda_{\hat{H}}$$
$$\lambda_E = b_{\hat{L}}\lambda_{\hat{L}} + b_{\hat{H}\tilde{L}}\lambda_{\hat{H}\tilde{L}} + (1-f_{\hat{H}\tilde{H}})b_{\hat{H}\tilde{H}}\lambda_{\hat{H}\tilde{H}} + f_{\hat{H}\tilde{H}}\lambda_{\hat{H}\tilde{H}} = b_{\hat{L}}\lambda_{\hat{L}} + b_{\hat{H}}\lambda_{\hat{H}} + f_{\hat{H}\tilde{H}}\beta(1-b_{\hat{H}})\lambda_{\hat{H}}$$

Hence we have

$$\begin{aligned}
\frac{\partial C_s(f_{\hat{H}\tilde{H}})}{\partial f_{\hat{H}\tilde{H}}} &= -[(1-\alpha)b_{\hat{H}} + \beta(1-b_{\hat{H}})]\lambda_{\hat{H}}(a_G + wQ_G) \\
&\quad + \beta(1-b_{\hat{H}})\lambda_{\hat{H}}[a_E + wQ_E(\lambda_E) + \lambda_E w\frac{\partial Q_E(\lambda_E)}{\partial\lambda_E}] \\
&= [(1-\alpha)b_{\hat{H}} + \beta(1-b_{\hat{H}})]\lambda_{\hat{H}}[(1-b_{\hat{H}\tilde{H}})(a_E + wQ_E(\lambda_E) + \lambda_E w\frac{\partial Q_E(\lambda_E)}{\partial\lambda_E}) - (a_G + wQ_G)] \\
&\leq [(1-\alpha)b_{\hat{H}} + \beta(1-b_{\hat{H}})]\lambda_{\hat{H}}[(1-b_{\hat{H}\tilde{H}})M_E(b_{\hat{L}}\lambda_{\hat{L}} + \lambda_{\hat{H}}) - (a_G + wQ_G)] \\
&\leq [(1-\alpha)b_{\hat{H}} + \beta(1-b_{\hat{H}})]\lambda_{\hat{H}}[(1-b_{\hat{H}})M_E(b_{\hat{L}}\lambda_{\hat{L}} + \lambda_{\hat{H}}) - (a_G + wQ_G)] \\
&\leq 0
\end{aligned}$$

Hence we have $f^*_{\hat{H}\tilde{H}} = 1$.

Now we have

$$C^*_s(\alpha,\beta) = \lambda_G w Q_G + \lambda_E w Q_E(\lambda_E) + \lambda_G a_G + \lambda_E a_E \tag{EC.42}$$

where

$$\begin{aligned}
\lambda_G &= \lambda_{\hat{L}\tilde{L}} + (1-f^*_{\hat{L}\tilde{H}})\lambda_{\hat{L}\tilde{H}} + (1-f^*_{\hat{H}\tilde{L}})\lambda_{\hat{H}\tilde{L}} \\
&= \lambda_{\hat{L}} - f^*_{\hat{L}\tilde{H}}[(1-\alpha)b_{\hat{L}} + \beta(1-b_{\hat{L}})]\lambda_{\hat{L}} + (1-f^*_{\hat{H}\tilde{L}})[\alpha b_{\hat{H}} + (1-\beta)(1-b_{\hat{H}})]\lambda_{\hat{H}} \\
\lambda_E &= b_{\hat{L}\tilde{L}}\lambda_{\hat{L}\tilde{L}} + (1-f^*_{\hat{L}\tilde{H}})b_{\hat{L}\tilde{H}}\lambda_{\hat{L}\tilde{H}} + f^*_{\hat{L}\tilde{H}}\lambda_{\hat{L}\tilde{H}} + (1-f^*_{\hat{H}\tilde{L}})b_{\hat{H}\tilde{L}}\lambda_{\hat{H}\tilde{L}} + f^*_{\hat{H}\tilde{L}}\lambda_{\hat{H}\tilde{L}} + \lambda_{\hat{H}\tilde{H}} \\
&= b_{\hat{L}}\lambda_{\hat{L}} + f^*_{\hat{L}\tilde{H}}\beta(1-b_{\hat{L}})\lambda_{\hat{L}} + b_{\hat{H}}\lambda_{\hat{H}} + f^*_{\hat{H}\tilde{L}}(1-\beta)(1-b_{\hat{H}})\lambda_{\hat{H}} + \beta(1-b_{\hat{H}})\lambda_{\hat{H}}
\end{aligned}$$

We have

$$\frac{\partial C^*_s(\alpha,\beta)}{\partial\alpha} = \frac{\partial C^*_s(\alpha,\beta,f^*_{\hat{L}\tilde{H}},f^*_{\hat{H}\tilde{L}})}{\partial\alpha} + \frac{\partial C^*_s(\alpha,\beta,f^*_{\hat{L}\tilde{H}},f^*_{\hat{H}\tilde{L}})}{\partial f^*_{\hat{L}\tilde{H}}}\frac{\partial f^*_{\hat{L}\tilde{H}}}{\partial\alpha} + \frac{\partial C^*_s(\alpha,\beta,f^*_{\hat{L}\tilde{H}},f^*_{\hat{H}\tilde{L}})}{\partial f^*_{\hat{H}\tilde{L}}}\frac{\partial f^*_{\hat{H}\tilde{L}}}{\partial\alpha} \tag{EC.43}$$

In $R^e_{0,1}, R^e_{1,1}, R^e_{0,0}$ and $R^e_{1,0}$, we have $\frac{\partial f^*_{\hat{L}\tilde{H}}}{\partial\alpha} = 0$ and $\frac{\partial f^*_{\hat{H}\tilde{L}}}{\partial\alpha} = 0$. In $R^e_{(0,1),1}$ and $R^e_{(0,1),0}$, we have $\frac{\partial C^*_s(\alpha,\beta,f^*_{\hat{L}\tilde{H}},f^*_{\hat{H}\tilde{L}})}{\partial f^*_{\hat{L}\tilde{H}}} = 0$ and $\frac{\partial f^*_{\hat{H}\tilde{L}}}{\partial\alpha} = 0$. In $R^e_{0,(0,1)}$ and $R^e_{1,(0,1)}$, we have $\frac{\partial f^*_{\hat{L}\tilde{H}}}{\partial\alpha} = 0$ and $\frac{\partial C^*_s(\alpha,\beta,f^*_{\hat{L}\tilde{H}},f^*_{\hat{H}\tilde{L}})}{\partial f^*_{\hat{H}\tilde{L}}} = 0$. Hence, we have

$$\frac{\partial C^*_s(\alpha,\beta)}{\partial\alpha} = \frac{\partial C^*_s(\alpha,\beta,f^*_{\hat{L}\tilde{H}},f^*_{\hat{H}\tilde{L}})}{\partial\alpha} = [f^*_{\hat{L}\tilde{H}}b_{\hat{L}}\lambda_{\hat{L}} + (1-f^*_{\hat{H}\tilde{L}})b_{\hat{H}}\lambda_{\hat{H}}](a_G + wQ_G) \geq 0 \tag{EC.44}$$

Similarly, we have

$$\frac{\partial C^*_s(\alpha,\beta)}{\partial\beta} = \frac{\partial C^*_s(\alpha,\beta,f^*_{\hat{L}\tilde{H}},f^*_{\hat{H}\tilde{L}})}{\partial\beta} + \frac{\partial C^*_s(\alpha,\beta,f^*_{\hat{L}\tilde{H}},f^*_{\hat{H}\tilde{L}})}{\partial f^*_{\hat{L}\tilde{H}}}\frac{\partial f^*_{\hat{L}\tilde{H}}}{\partial\beta} + \frac{\partial C^*_s(\alpha,\beta,f^*_{\hat{L}\tilde{H}},f^*_{\hat{H}\tilde{L}})}{\partial f^*_{\hat{H}\tilde{L}}}\frac{\partial f^*_{\hat{H}\tilde{L}}}{\partial\beta} \tag{EC.45}$$

In $R^e_{0,1}, R^e_{1,1}, R^e_{0,0}$ and $R^e_{1,0}$, we have $\frac{\partial f^*_{\hat{L}\tilde{H}}}{\partial\beta} = 0$ and $\frac{\partial f^*_{\hat{H}\tilde{L}}}{\partial\beta} = 0$. In $R^e_{(0,1),1}$ and $R^e_{(0,1),0}$, we have $\frac{\partial C^*_s(\alpha,\beta,f^*_{\hat{L}\tilde{H}},f^*_{\hat{H}\tilde{L}})}{\partial f^*_{\hat{L}\tilde{H}}} = 0$ and $\frac{\partial f^*_{\hat{H}\tilde{L}}}{\partial\beta} = 0$. In $R^e_{0,(0,1)}$ and $R^e_{1,(0,1)}$, we have $\frac{\partial f^*_{\hat{L}\tilde{H}}}{\partial\beta} = 0$ and $\frac{\partial C^*_s(\alpha,\beta,f^*_{\hat{L}\tilde{H}},f^*_{\hat{H}\tilde{L}})}{\partial f^*_{\hat{H}\tilde{L}}} = 0$. Hence, we have

$$\begin{aligned}
\frac{\partial C^*_s(\alpha,\beta)}{\partial\beta} &= \frac{\partial C^*_s(\alpha,\beta,f^*_{\hat{L}\tilde{H}},f^*_{\hat{H}\tilde{L}})}{\partial\beta} \\
&= [f^*_{\hat{L}\tilde{H}}b_{\hat{L}}\lambda_{\hat{L}} + (1-f^*_{\hat{H}\tilde{L}})b_{\hat{H}}\lambda_{\hat{H}}][a_E + wQ_E(\lambda_E) + \lambda_E w\frac{\partial Q_E(\lambda_E)}{\partial\lambda_E} - a_G - wQ_G] \\
&\geq 0
\end{aligned} \tag{EC.46}$$

When $\alpha+\beta=1$, we have $b_{\hat{L}\tilde{L}} = b_{\hat{L}\tilde{H}} = b_{\hat{L}}$ and $b_{\hat{H}\tilde{L}} = b_{\hat{H}\tilde{H}} = b_{\hat{H}}$, and therefore $C^*_s(\alpha,\beta) = C_s(f_{\hat{L}} = 0, f_{\hat{H}} = 1)$. In addition, we have $\frac{\partial C^*_s(\alpha,\beta)}{\partial\alpha} \geq 0$ and $\frac{\partial C^*_s(\alpha,\beta)}{\partial\beta} \geq 0$. Hence, we have $C^*_s(\alpha,\beta) \leq C_s(f_{\hat{L}} = 0, f_{\hat{H}} = 1), \forall\ \alpha,\beta$ s.t. $\alpha\geq 0, \beta\geq 0, \alpha+\beta\leq 1$. $\square$

Proof of Proposition 7. Given $f_{\hat{L}\tilde{L}}=0$ and $f_{\hat{H}\tilde{H}}=1$, we have

$$\begin{aligned}\lambda_G &= \lambda_{\hat{L}\tilde{L}} + (1-f_{\hat{L}\tilde{H}})\lambda_{\hat{L}\tilde{H}} + (1-f_{\hat{H}\tilde{L}})\lambda_{\hat{H}\tilde{L}} \\ &= \lambda_{\hat{L}} - f_{\hat{L}\tilde{H}}[(1-\alpha)b_{\hat{L}} + \beta(1-b_{\hat{L}})]\lambda_{\hat{L}} + (1-f_{\hat{H}\tilde{L}})[\alpha b_{\hat{H}} + (1-\beta)(1-b_{\hat{H}})]\lambda_{\hat{H}} \\ \lambda_E &= b_{\hat{L}\tilde{L}}\lambda_{\hat{L}\tilde{L}} + (1-f_{\hat{L}\tilde{H}})b_{\hat{L}\tilde{H}}\lambda_{\hat{L}\tilde{H}} + f_{\hat{L}\tilde{H}}\lambda_{\hat{L}\tilde{H}} + (1-f_{\hat{H}\tilde{L}})b_{\hat{H}\tilde{L}}\lambda_{\hat{H}\tilde{L}} + f_{\hat{H}\tilde{L}}\lambda_{\hat{H}\tilde{L}} + \lambda_{\hat{H}\tilde{H}} \\ &= b_{\hat{L}}\lambda_{\hat{L}} + f_{\hat{L}\tilde{H}}\beta(1-b_{\hat{L}})\lambda_{\hat{L}} + b_{\hat{H}}\lambda_{\hat{H}} + f_{\hat{H}\tilde{L}}(1-\beta)(1-b_{\hat{H}})\lambda_{\hat{H}} + \beta(1-b_{\hat{H}})\lambda_{\hat{H}}\end{aligned}$$

The social cost is

$$C_s(\alpha,\beta) = \lambda_G w Q_G + \lambda_E w Q_E(\lambda_E) + \lambda_G a_G + \lambda_E a_E \tag{EC.47}$$

We then have

$$\begin{aligned}\frac{\partial C_s(\alpha,\beta)}{\partial f_{\hat{L}\tilde{H}}} &= -[(1-\alpha)b_{\hat{L}} + \beta(1-b_{\hat{L}})]\lambda_{\hat{L}}(a_G + wQ_G) \\ &\quad + \beta(1-b_{\hat{L}})\lambda_{\hat{L}}[a_E + wQ_E(\lambda_E) + \lambda_E w\frac{\partial Q_E(\lambda_E)}{\partial \lambda_E}]\end{aligned} \tag{EC.48}$$

and

$$\begin{aligned}\frac{\partial C_s(\alpha,\beta)}{\partial f_{\hat{H}\tilde{L}}} &= -[\alpha b_{\hat{H}} + (1-\beta)(1-b_{\hat{H}})]\lambda_{\hat{H}}(a_G + wQ_G) \\ &\quad + (1-\beta)(1-b_{\hat{H}})\lambda_{\hat{H}}[a_E + wQ_E(\lambda_E) + \lambda_E w\frac{\partial Q_E(\lambda_E)}{\partial \lambda_E}]\end{aligned} \tag{EC.49}$$

On the other hand, we have the potential function

$$\Phi(\alpha,\beta) = \int_0^{\lambda_G} wQ_G dx + \int_0^{\lambda_E} wQ_E(x)dx + \lambda_G p_G + \lambda_E p_E \tag{EC.50}$$

where the equilibrium patient flow under p_G and p_E is given by

$$\frac{\partial \Phi(\alpha,\beta)}{\partial f_{\hat{L}\tilde{H}}} = -[(1-\alpha)b_{\hat{L}} + \beta(1-b_{\hat{L}})]\lambda_{\hat{L}}(a_G + wQ_G) + \beta(1-b_{\hat{L}})\lambda_{\hat{L}}[p_E + wQ_E(\lambda_E)] \tag{EC.51}$$

and

$$\frac{\partial \Phi(\alpha,\beta)}{\partial f_{\hat{H}\tilde{L}}} = -[\alpha b_{\hat{H}} + (1-\beta)(1-b_{\hat{H}})]\lambda_{\hat{H}}(a_G + wQ_G) + (1-\beta)(1-b_{\hat{H}})\lambda_{\hat{H}}[p_E + wQ_E(\lambda_E)] \tag{EC.52}$$

Comparing EC.48 with EC.51 and EC.49 with EC.52, we can see that optimal patient flow can be induced by setting $p_G = a_G$ and $p_E = a_E + \frac{\partial Q_E(\lambda_E)}{\partial \lambda_E}$. Hence, we have $\tilde{p}_G^*(\alpha,\beta) = a_G = \hat{p}_G^*$. In addition, when we have $f_{\hat{L}\tilde{H}}^* \in (0,1)$ or $f_{\hat{H}\tilde{L}}^* \in (0,1)$, $p_E = a_E + \frac{\partial Q_E(\lambda_E)}{\partial \lambda_E}$ is the minimum associated expected ED fee for $\tilde{p}_G^*(\alpha,\beta) = a_G$, and therefore we have $\tilde{p}_E^*(\alpha,\beta) = a_E + \frac{\partial Q_E(\lambda_E)}{\partial \lambda_E} > \hat{p}_E^*$. □

References

Roughgarden T. 2007. Routing games. Nisan N, Roughgarden T, Tardos É, Vazirani V. *Algorithmic Game Theory* (Cambridge University Press, New York) 461–486.